UNIVERSITÀ DEGLI STUDI DI BARI
Dipartimento Interateneo di Fisica "M. Merlin"

DOTTORATO DI RICERCA IN FISICA
Ciclo XXXIV
Settore scientifico disciplinare Fis/01

Search for Higgs Production via Vector Boson Fusion in the four-lepton final state by using CMS run II data at 13 TeV

Dottoranda:

Walaa Mohamed Mohamed Abdelaziz Elmetenawee

Supervisori:

Prof. **Nicola De Filippis**
Prof. **Marcello Abbrescia**

Coordinatore:

Ch.mo Prof. **Giuseppe Gonnella**

ESAME FINALE 2022

Abstract

This thesis presents a standalone measurement of the Higgs boson production cross-section in the Vector Boson Fusion Vector Boson Fusion (VBF) production mode performed in the $H \rightarrow ZZ \rightarrow 4l(l = e,\ \mu)$ channel. The study is done using datasets corresponding to proton-proton collision data produced at the LHC at $\sqrt{s} = 13$ TeV and collected by the CMS detector, equivalent to an integrated luminosity of 137 fb^{-1}. The artificial neural network technique is used in order to perform the multivariate analysis for two orthogonal jet-based categories defined in this analysis. Statistical shape analysis developed for the two jet-based categories is done by combining the discriminant distribution from each category. The Higgs VBF signal strength modifier and the significance of the present analysis have been estimated. The present analysis has enough significance for the VBF Higgs production evidence. A projection for future luminosities is also presented, and accordingly, it is expected to have enough significance for the VBF Higgs production observation (5σ) with an integrated luminosity of 350 fb^{-1}.

Contents

List of Figures

List of Tables

Introduction

One of the biggest mysteries of the Standard Model (SM) was to understand the mechanism responsible for breaking the electroweak symmetry and generates the masses of the elementary particles. According to the theory, the elementary particles have been initially treated as non-massive objects to keep the certain symmetry conditions needed since physics must be invariant under frame variation. However, around 1964, an idea for spontaneously generating the mass of the particles without breaking their symmetry was conceived in theory. [1–3].

In 2012, the last missing piece on the SM basis had been found by the discovery of a new boson, the so-called "Higgs boson" at the LHC at CERN by the ATLAS and CMS collaborations [4, 5] which represents, therefore, a significant achievement in the particle physics field.

After the Higgs discovery, the properties of the Higgs boson have been measured in many of the accessible final states. From the combination of the ATLAS and CMS measurements [6], the Higgs boson mass has been measured to be 125.09 $\pm$ 0.21 (stat) $\pm$0.11 (syst) GeV. Several results from both experiments demonstrated that its measured properties, including its spin, coupling strengths to fermions and bosons, and the CP properties, are compatible with the SM expectations.

With increasing the amount of data collected from the LHC, the properties of the Higgs boson can be measured with better precision, and rarer decay modes become accessible. Such measurements are interesting because any deviation from the theory's prediction might be a hint for a new physics beyond the SM. The Vector Boson Fusion (VBF) production mode is a key process for precision measurements of Higgs properties since it is a clean channel with very distinctive kinematics. In addition, it is followed by two energetic jets separated by a η gap, which can be used as a tagging input. Such features provide a good scenario for measurements of Higgs couplings [7]. The VBF channel is a direct probe of the coupling between vector bosons and the Higgs boson and, hence, directly probes the electroweak sector of the SM.

This thesis presents a dedicated measurement of the Higgs boson production cross-section in the VBF production mode in the $H \rightarrow ZZ \rightarrow 4l(l = e,\ \mu)$ channel. The dataset used in the analysis corresponds to proton-proton collision data produced at the LHC at a center-of-mass energy of 13 TeV and collected by the CMS detector equivalent to an integrated luminosity of $137fb^{-1}$. The present analysis relies essentially on the reconstruction, identification, and isolation of leptons and jets. The selections performed in this analysis are designed to reconstruct a final state with four charged leptons ($4\mu, 4e$ or $2e2\mu$) and at least two jets. The four-lepton system was designed to reconstruct the Higgs boson and is synchronized to the one used to discover the Higgs in the $ZZ \rightarrow 4l$ channel.

To enhance and discriminate the VBF signal, I used Artificial Neural Network (ANN) trained on Monte Carlo (MC) events for signal and background, studied a large number of configurations of the network parameters, and selected the one providing the best efficiency times purity. The analysis proposes the usage of the 3^{rd} jet when available in the event; therefore, a new sample for VBF Higgs production associated with three hard jets at NLO was generated (private production). In order to optimize the analysis and adequately handle the usage of the third jet in the events, two jet-based categories have been defined. For each category, a NN discriminant was developed.

The thesis is structured as follows:

- Chapter 1 introduces the SM of particle physics, with an emphasis on the Higgs mechanism. Moreover, an overview of the various decays and production modes of the Higgs boson is also discussed, motivating the VBF production mode used in this analysis as one of the most promising and interesting signals. Finally, the most important measurements of the Higgs boson properties are also summarized in that chapter.

- Chapter 2 is dedicated to the experimental setup used to obtain the data for this analysis. The chapter starts with a brief overview of the LHC accelerator and its experiments, followed by a more elaborate description of the CMS experiment, its detectors and its trigger system. Then the second part of this chapter covers the reconstruction and identification of the physics objects used in this analysis and their performance. It also reviews the Monte Carlo event simulation used in this thesis.

- Chapter 3 presents a study I have done to measure the CMS tracking performance, which is crucial for the overall performance of the CMS experiment. The study has been done using the tag-and-probe method exploiting the $Z \to \mu^+\mu^-$ resonance.

- Chapter 4 presents a study I have done to evaluate the effects of the electronic threshold on the performance of the RPC system of the CMS experiment. This study aims to optimize the RPC performance, and hence the effects on the RPC efficiency, cluster size and detector intrinsic noise rate of variations of the electronics has been discussed.

- Chapter 5 represents the building blocks required to perform the analysis. It describes the dataset used, the signal samples, and the expected background processes. The second part of the chapter is dedicated to a detailed description of the selection criteria applied to choose the physics objects used in the thesis.

- Chapter 6 gives an overview of the analysis strategy followed to in this work, starting with the event selections that applied to define the VBF signal region, then it described the background estimation, including the data-driven estimation for the Z+X background.

- Chapter 7 introduces machine learning concepts and the strategy of using ANN to search for the SM VBF Higgs boson signal. The Chapter starts with a description of the general architecture of the ANN, the learning process, the training strategy. Then, it presents the results of the performance of the trained ANN for each year used in the analysis. Finally, it ends with a comparison between data and simulation for the output discriminant.

- Chapter 8 introduces the sources of systematic uncertainties affecting the measurements. In addition, it describes the statistical procedure used to evaluate the signal's presence. The last part of the chapter presents the final yield and distributions and the statistical interpretation of the results.

Chapter 1

The Standard Model and the Higgs Boson

The endeavor of particle physics lies in the observation, formulation, and validation of the rules that describe the fundamental interactions between elementary particles and their properties. Our current understanding of the elementary particles is summarized in the so-called the SM of particle physics. The SM gathers quantum field theories which describe the elementary particles building the ordinary matter and the interaction between them via three mechanisms: the strong nuclear, the weak nuclear, and the electromagnetic interactions. The gravitational interaction is not included, but it is much weaker and is not expected to contribute to the physical processes currently investigated in high-energy physics. An overview of the SM and a brief description of the Higgs boson mechanism are given in this chapter to motivate the Higgs characterization effort and the analysis that is targeting the VBF Higgs production observation.

1.1 Fundamental particles and interactions

The elementary particles are the ones that do not have an internal structure; they are classified as fermions and bosons. The former have S = 1/2, and they are fundamental constituents of the matter; in turn, they are classified as quarks and leptons. The latter have S = 1 and mediate the interactions between fermions. An overview of the particle content of the SM and their properties are given in Fig. 1.1.

Twelve fermions are experimentally observed and included in the SM. Each of them has an "antiparticle", e.g. a particle with identical mass and opposite quantum numbers. Omitting their antiparticles, the known electrically charged leptons are the electron e, the muon μ, and the tau lepton τ, with Q = −1; each is paired to a neutrino $(\nu_e, \nu_\mu, \nu_\tau)$, electrically neutral, to form a "generation". Their masses span a wide range: the electron has mass m_e = 511 MeV, and it is the lightest of the charged leptons; the muon and the tau lepton have masses about 200 and 3500 times larger; neutrinos are massless in the classical SM formulation. However, the observation of neutrino oscillations implies that also neutrinos are massive [9, 10] ($m_\nu < 2$ eV); to account for it, a modification is introduced in the SM formalism through the PMNS matrix [11].

Each quark carries a quantum number called "flavor", subjected to the electroweak interactions;

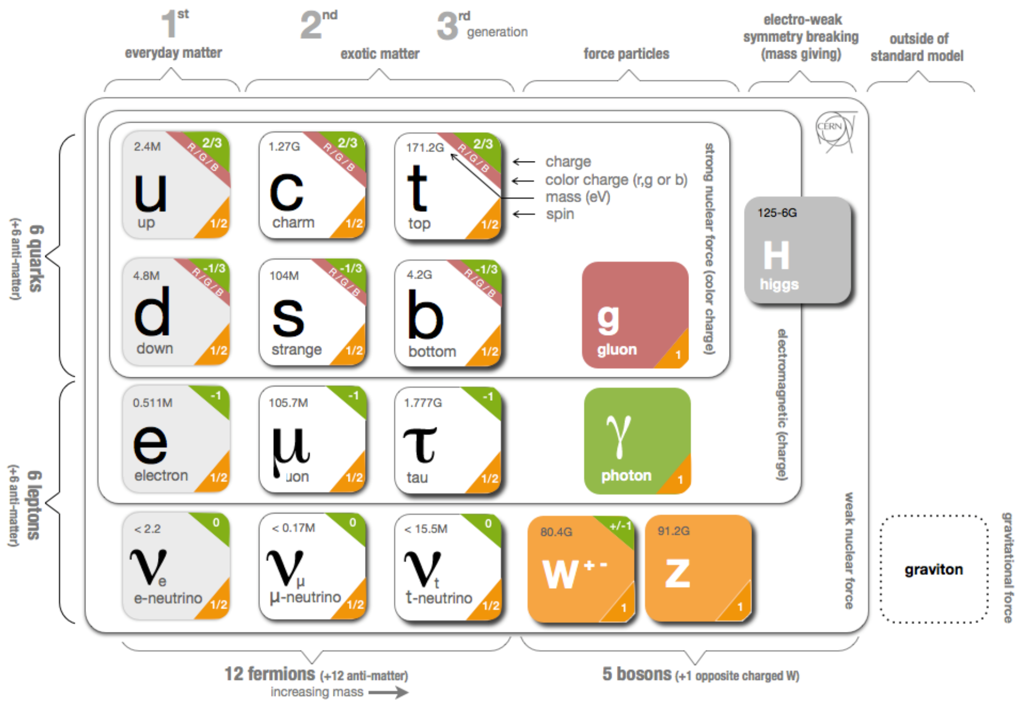


Figure 1.1: Diagram representing the particle content of the Standard Model [8]. The graviton, represented outside of the Standard Model in this figure, is not observed.

like leptons, the quark flavors are paired in three generations: up (u) and down (d), charm (c) and strange (s), top (t) and beauty (b). Within each pair, the first quark has Q $= 2/3$ and the second Q $= -1/3$. In addition to the other quantum numbers, quarks carry a "color", denoted as $q_i = 1, 2, 3$, on which the strong interaction operates by the rules described by the QCD theory. Quarks are only experimentally observed as color-neutral bonds, called "hadrons"; this QCD property is known as "confinement". Like charged leptons, quarks also have masses that differ by several orders of magnitude: the heaviest quark, the top, has a $m_\tau = 172.9$ GeV mass; thus, it has an extremely short lifetime and, before a hadron can be formed, it decays through weak interaction into a W boson and a lighter quark [10]. However, quarks are confined in hadrons; they are "asymptotically free" particles [12] in high energy collisions, i.e., the strong coupling becomes weaker when the momentum transfer is large. This feature allows the fundamental interactions between them to be investigated in High Energy Physics with proton-proton colliders such as the LHC.

The SM describes the fundamental interactions between the particles as a quantum field theory with an exchange of force-mediating particles. Three of the four known forces are incorporated in the SM, and their interactions and kinematics are described by mathematical constructions called Lagrangian density. The SM Lagrangian is invariant under Lorentz transformation that implies the conservation of energy and momentum.

In the SM, the description of the fundamental interactions is based on the local gauge symmetry group [13, 14]

$$\underbrace{SU(3)_C}_{QCD} \times \underbrace{SU(2)_L \times U(1)_Y}_{EW} \tag{1.1}$$

$SU(3)_C$ describes the strong interactions and the index "C" refers to the color charge, which is conserved under $SU(3)$ symmetry. The gauge sector of the SM contains gluons, which mediate the strong interaction and correspond to the generators of the $SU(3)_C$ group. $SU(N)$ groups with dimension N require $(N^2 - 1)$ generators; therefore, $SU(3)_C$ has eight generators corresponding to eight gluons. Gluons carry color quantum numbers for red, green, and blue; therefore, they interact with each other and with colored particles similar to quarks. Under $SU(3)_C$, quarks are color triplets while leptons are color singlets; therefore, quarks carry a color index ranging between one and three, whereas leptons do not take part in strong interactions.

$SU(2)_L \times U(1)_Y$ describes the unification of the weak and electromagnetic forces in the so-called non-abelian gauge group of the electroweak interaction. The "L" subscript refers to the left chiral nature of the $SU(2)_L$ coupling, and "Y" in $U(1)_Y$ refers to the weak hypercharge which is carried by the gauge field B_μ and is relevant to the electric charge Q and the weak isospin T_3 as

$$Y = 2(Q - T_3) \tag{1.2}$$

The three gauge fields W_μ^1, W_μ^2, and W_μ^3 are associated to $SU(2)_L$ with three generators that can be written as half the Pauli matrices:

$$T_1 = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ T_2 = \frac{1}{2}\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix},\ T_3 = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{1.3}$$

The generators T_a obeying the commutator rule:

$$[T_a, T_b] = i \sum_{c=1}^{3} \epsilon_{abc} T_c \tag{1.4}$$

where ϵ_{abc} is a fully anti-symmetric tensor.

In the SM, fermions (quarks and leptons) are represented by matter fields or Dirac (spinor) fields, denoted by ψ, that is split into left and right-handed parts (ψ^L, ψ^R). Left-handed charged leptons and the corresponding neutrino spinors are grouped in weak isospin doublets. The same in the case of quarks, left-handed up-type quark fields are grouped with the corresponding down-type quark fields.

$$\psi_q^L = \begin{pmatrix} \psi_{u_q}^L \\ \psi_{d_q}^L \end{pmatrix}, \psi_{u_q}^R, \psi_{d_q}^R \tag{1.5}$$

Similarly for leptons,

$$\psi_l^L = \begin{pmatrix} \psi_{\nu_l}^L \\ \psi_l^L \end{pmatrix}, \psi_l^R = l_R \tag{1.6}$$

where u,d represent up-type and down-type quarks, ψ_l^L is the left-handed lepton field, l is an index over the three lepton generations, and ψ_l^R is the right-handed field. One should notice that in the SM there are no right-handed neutrinos.

The matter fields transform linearly under the gauge group. In order to preserve the local gauge invariance, all the field space-time derivatives are promoted to covariant derivatives: $\partial\mu \rightarrow D_\mu$. Given the SM gauge structure, the covariant derivative will be:

$$D_\mu = \partial\mu + ig'YB_\mu(x) + igW_\mu^a(x)T^a + ig_S G_\mu^a(x)t^a \tag{1.7}$$

where g', g are the coupling of the hypercharge field B_μ and the electroweak interactions, mediated by the three W_μ^a fields; g_S is the coupling of the strong interaction, mediated by the eight G_a^μ fields. The vector boson fields are functions of the space-time (the 4-vector x), as explicitly written in the above formula. The Y, T^a and t^a are the generators of the symmetry group in the representation required by the field. When substituted in quadratic terms, the covariant derivative produces the interaction vertices between fermions and vector bosons. This description of the interaction vertices, also called minimal coupling substitution, is predictive when used in perturbation theory.

The invariant kinetic terms for the gauge bosons are also part of the Lagrangian, in the form:

$$\mathcal{L}_{kin} = -\frac{1}{4}G_{\mu\nu}^a G^{a\mu\nu} - \frac{1}{4}W_{\mu\nu}^a W^{a\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \tag{1.8}$$

where the field strength tensors $G_{\mu\nu}^a, W_{\mu\nu}^a$ and $B^{\mu\nu}$ appear. The field strength tensor is defined, for the weak gauge fields, as:

$$W_{\mu\nu}^a = \partial\mu W_\nu^a + \partial\nu W_\mu^a - igf^{abc}W_\mu^b W_\nu^b \tag{1.9}$$

The third term only appears if the group corresponding to the gauge field is non-abelian; g is the coupling that is also found in the covariant derivative. In the SM, the $SU(2)_L$ and $SU(3)_c$ groups are non-abelian, while the hypercharge group $U(1)_Y$ is abelian. The third term generates both cubic and quartic self-interactions among the gauge fields.

The gauge bosons have no mass terms, as those would not be invariant under the gauge symmetry. A possible solution to give a mass to the gauge bosons is the Higgs mechanism discussed in the next section.

1.2 The Higgs mechanism

The Brout-Englert-Higgs (BEH) mechanism, also known as electroweak symmetry breaking (EWSB), was introduced to generate the gauge boson masses and explain the fermion masses, which were not accounted for in the electroweak gauge formalism. It was proposed in 1964 independently by F. Englert and R. Brout [1], by P. Higgs [3], and by G. Guralnik, C. R. Hagen, and T. Kibble [15]. The BEH mechanism represents a symmetry-breaking process that provides the SM with a renormalizable [16] and gauge-invariant Lagrangian that accounts for the mass of the gauge bosons. In the simplest case, Spontaneous Symmetry Breaking (SSB) requires the presence of a scalar field with a positive vacuum expectation value. In the SM case, the SSB is realized via a complex scalar field which spontaneously breaks the $SU(2)_L \times U(1)_Y$ gauge symmetry [1–3]. The $SU(3)_C$ symmetry remains unbroken.

The additional Lagrangian terms due to the scalar Higgs field is given by

$$\mathcal{L}_{Higgs} = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi) \tag{1.10}$$

To spontaneously break the symmetry, consider a complex scalar field $\Phi(x)$ which behaves as doublet under $SU(2)_L$ and has a hypercharge value of 1,

$$\Phi(x) = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \tag{1.11}$$

Renormalizability and $SU(2)_L \times U(1)_Y$ invariance claim the Higgs potential $V(\Phi)$ to take the form:

$$V(\Phi) = V(\Phi^\dagger \Phi) = -\mu^2 \Phi^\dagger \Phi + \lambda(\Phi^\dagger \Phi)^2 \tag{1.12}$$

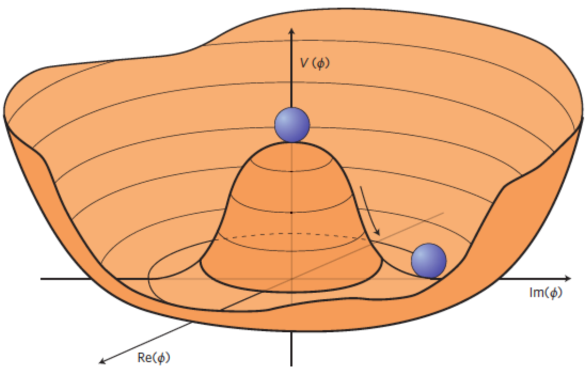


Figure 1.2: Illustration of the Higgs potential $V(\Phi)$ for the complex scalar field. Whereas the potential itself is symmetric, choosing any of the points at the minimum of the potential effectively breaking the symmetry for non-vanishing fields Φ. [17]

The peculiarity of the potential $V(\Phi)$ is its symmetric shape that provides asymmetric ground states, as shown in Fig. 1.2. The symmetry is spontaneously broken when the minimum of the potential $V(\Phi)$, also called vacuum expectation value, is obtained for non-vanishing scalar fields Φ in the configurations where the constants μ^2 and λ are positive. The Φ field, with appropriate gauge fixing, can be rewritten as

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ \nu + H(x) \end{pmatrix} \quad with \;\; \nu = \left(\frac{\mu^2}{2\lambda}\right)^{1/2} \tag{1.13}$$

In this way, the symmetry is broken, generating a mass for the corresponding gauge boson. The three gauge bosons acquire a mass, whereas the $U(1)_{em}$ symmetry persists unbroken, the photon remains massless. The covariant derivative term in the Lagrangian equation can be rewritten as:

$$\frac{1}{2}\partial_\mu H(x)\partial^\mu H(x) + \frac{1}{8}(gW_\mu^3 - g'B_\mu)^2 \cdot (\nu + H(x))^2 + \frac{1}{8}g^2(W_\mu^1 - iW_\mu^2)(W^{1\mu} + iW^{2\mu}) \cdot (\nu + H(x))^2 \tag{1.14}$$

The mass of the gauge bosons arise through the combination of the fields $W_\mu^{1,2}$ (for the bosons $W^\pm$) and of the fields Y_μ and W_μ^3 (for the bosons Z and γ -the photon) through the expressions:

$$W_\mu^\pm = \frac{W_\mu^1 \pm iW_\mu^2}{\sqrt{2}}, \qquad M_W = \frac{\nu g}{2} \tag{1.15}$$

$$Z_\mu = \frac{gW_\mu^3 - g'B_\mu}{\sqrt{g^2 + g'^2}}, \qquad M_Z = \frac{\nu\sqrt{g^2 + g'^2}}{2} \tag{1.16}$$

$$A_\mu = \frac{g'W_\mu^3 + gB_\mu}{\sqrt{g^2 + g'^2}}, \qquad M_A = 0 \tag{1.17}$$

The physical Higgs field, $H(x)$, describes a scalar particle with mass $m_H = \sqrt{2\lambda} \cdot \nu$ which can be derived from the potential term $V(\Phi)$. The mass of the Higgs is a free parameter that depends on the λ introduced by the Higgs potential while the vacuum expectation value can be extracted from Fermi constant G_F to be $\nu = 246$ GeV.

Yukawa interaction

Thus far, the SSB of the local gauge invariance of $SU(2)_L \times U(1)_Y$ gives rise to the mass terms of the weak bosons through their interaction with the Higgs field. The same field allows mass terms proportional to the vacuum expectation value for the fermions while keeping the local gauge invariance [18, 19] through the so-called Yukawa interaction. The mass terms for the fermions take the form $m\left(\bar{\psi}_R\psi_L + \bar{\psi}_L\psi_R\right)$ which is not invariant under $SU(2)_L$. Gauge invariant mass terms are introduced via the interaction of the fermions with the Higgs field.

Accounting for the first lepton generation, their interaction is described by the Yukawa Lagrangian

$$\mathcal{L}_{Yukawa} = -y_e(\bar{\psi}_R\Phi^\dagger\psi_L + \bar{\psi}_L\Phi\psi_R) \tag{1.18}$$

where y_e is the coupling constant between an electron and a Higgs boson. By injecting the scalar doublet with parametrization Eq. 1.13, one has

$$\mathcal{L}_{Yukawa} = -y_e\frac{\nu + H(x)}{\sqrt{2}}\left(\bar{e}_Re_L + \bar{e}_Le_R\right) \tag{1.19}$$

re-grouping the electron fields as $\bar{e} = (\bar{e}_R, \bar{e}_L), e = (e_R, e_L)$, the Lagrangian can be written more conveniently as

$$\mathcal{L}_{Yukawa} = -y_e\frac{\nu}{\sqrt{2}}\bar{e}e - y_e\frac{H(x)}{\sqrt{2}}\bar{e}e \tag{1.20}$$

where the mass of the electron $m_e = y_e \nu/\sqrt{2}$ and the interaction term between the electron and the Higgs boson arise. Likewise, we can add the mass term for the other charged leptons with coupling proportional to the lepton mass.

The Higgs boson gives mass also to both the down type and up type quarks with Yukawa interactions. The mass terms for the down type quarks are analogous to the lepton masses, while for up type quark, the conjugated field

$$\widetilde{\Phi} = \frac{1}{\sqrt{2}} \begin{pmatrix} \nu + H(x)) \\ 0 \end{pmatrix} \tag{1.21}$$

is used in the mass terms which take the form:

$$\mathcal{L}_{Yukawa} = -y_u(\bar{u}_R \widetilde{\Phi}^\dagger q_L + \bar{q}_L \widetilde{\Phi} u_R) = y_u \frac{\nu + H(x)}{\sqrt{2}} (\bar{u}_R u_L + \bar{u}_L u_R) \tag{1.22}$$

As a result, the Higgs mechanism allows for both bosons and fermions to acquire masses by introducing the Higgs boson scalar field. The interaction of the Higgs boson with fermions and gauge bosons is related to the gauge couplings and the masses: as for fermions, the interaction is proportional to their mass, while the interaction with gauge bosons relies quadratically on their mass. As a consequence, the Higgs boson decays preferentially in the heaviest kinematically accessible particles.

1.3 Phenomenology of the Higgs boson at the LHC

On the 4^{th} July 2012, the ATLAS and CMS experiments at the LHC at ATLAS announced the observation of a new boson in the mass region of around 125 GeV compatible with the one predicted by the SM [4, 5]. The double confirmation of the Higgs boson led to the awarding of the 2013's Nobel Prize in Physics to François Englert and Peter Higgs for the discovery of the mechanism that generates the mass of the particles and for predicting the existence of the neutral massive scalar boson.

Once the Higgs boson mass value is assumed, predictions for its production and decay rates are possible. Four main Higgs boson production mechanisms are expected to occur at the LHC at the center-of-mass energies currently reached (up to 13 TeV). The leading Feynman diagrams are reported in Fig. 1.3. Their expected cross-sections for a Higgs boson of mass of 125 GeV are shown in Fig. 1.4 (a).

- **The gluon-gluon fusion (ggF)** is the dominant production mode over the entire mass spectrum at the LHC with a cross-section of 49 pb at $\sqrt{s}$ = 13 TeV, which approximately represent $\sim 85\%$ of the total. The two gluons of the protons interact to produce a Higgs boson via an intermediate heavy quark loop; the loop is necessary since the Higgs boson does not couple to the massless gluons. Being the Higgs boson couplings to fermions proportional to their mass, the main contribution to the SM amplitude comes from the top quark loop. It, therefore, is adequate to probe the top Yukawa coupling.

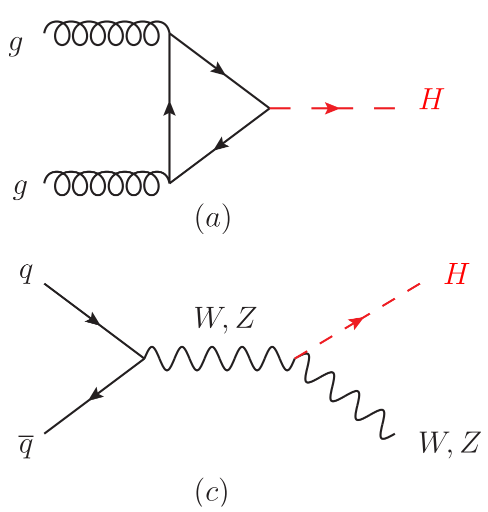


Figure 1.3: The Feynman diagrams for the Main Leading Order Higgs boson production modes in proton-proton collisions [20]; (a) gluon fusion, (b) Vector-boson fusion, (c) associated production with a gauge boson(Higgs-strahlung), (d) associated production with a pair of top quarks, their cross sections at the LHC are summarised in Fig. 1.4 (a)

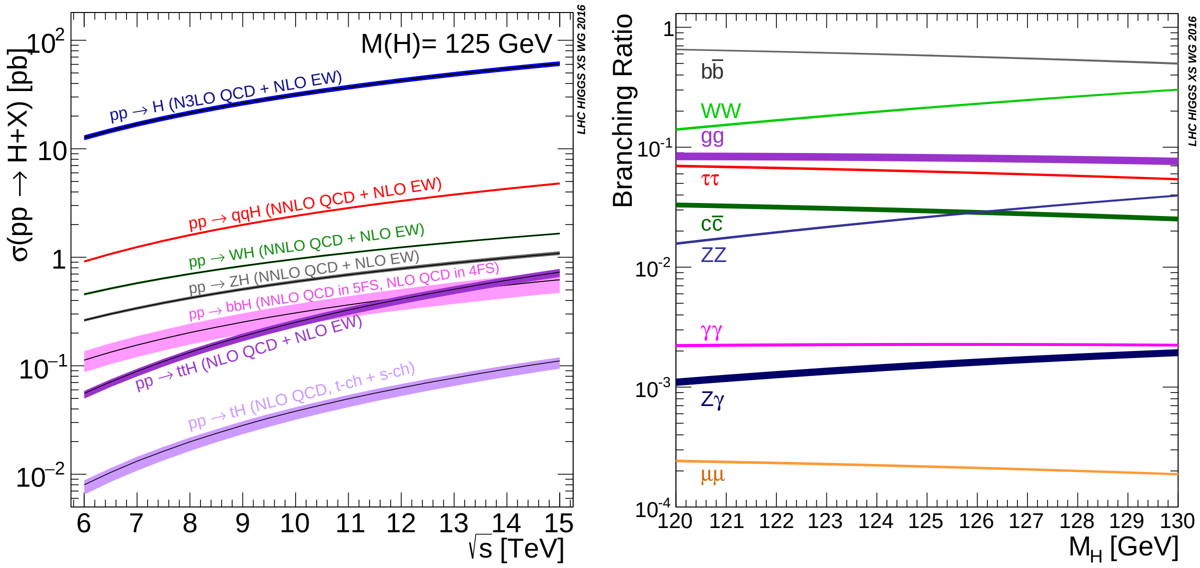
- **The vector boson fusion (VBF)** process has a cross-section of 3.8 pb at $\sqrt{s} = 13$ TeV, about ten times smaller than ggH. The Higgs boson is produced from the fusion of two W or two Z bosons which radiated from two quarks. The coupling of the Higgs boson to the vector bosons can be measured with this process. The outgoing quarks from the interaction proceed in the forward region, leaving a distinct signature of two energetic forward jets separated by a large pseudorapidity gap.

- **The Higgs-Strahlung (VH)** is the third most frequent mechanism, with a slightly smaller cross-section compared to the VBF. In this process, a quark interacts with an antiquark to produce a massive vector boson which radiates a Higgs boson. As in the VBF case, the VH process is instrumental in probing the Higgs boson coupling to vector bosons. At the LHC, although the cross-section of WH is approximately two times higher than the ZH, its precision in the reconstruction is limited due to the decays of the W boson to neutrinos, which are undetectable.

- **Top quark associated production (ttH)** is the last of the four main production modes, and its contribution to the total cross-section is about 1%. The Higgs boson is produced in association with a top quark and an anti-top quark which, in turn, originated from the gluon splitting. Unlike in the ggH case, there are no loops involved in these processes, allowing direct measurement of Higgs coupling to the top quark.

The most accurate predictions for the production cross-sections in each mode are reported in Fig. 1.4(a). The accuracy level in perturbation theory for the corrections due to the strong interactions, computed in perturbative QCD, and electroweak corrections is also reported in Fig. 1.4(a).

With a lifetime of 1.56×10^{-22} s, the Higgs boson decays immediately and it is studied via its decay products. It can decay to a variety of final states which can be divided in tree-level decays, such as ZZ, WW, $\tau^+\tau^-$, $b\bar{b}$, $\mu^+\mu^-$ and $c\bar{c}$, and via virtual loop to a pair of gluons or a pair of photons; as



(a) **(b)**

Figure 1.4: (a) Cross-sections of the main Higgs boson production modes as a function of the centre-of-mass energy for a mass of $m_H = 125$ GeV. (b) Branching ratios of the main decay modes of the Higgs boson as a function of the Higgs boson mass [21].

illustrated in Fig.1.5. The branching ratios of these decay channels depend on the Higgs mass, as reported in Fig.1.4(b).

Interestingly, even if the $H \to ZZ$ and $H \to \gamma\gamma$ have a lower branching ratio than $H \to WW$, they were the main drivers of the Higgs boson discovery in 2012. The former benefits from a high signal-to-background ratio and an excellent invariant mass resolution in leptonic decays; the latter profits from very high resolution in the reconstruction of the invariant diphoton mass.

1.4 The Higgs VBF Production Mode

Among the various production mechanisms of the Higgs boson, VBF is undoubtedly considered one of the most vitally important and promising signals. There are some special features of this production mode. It has the largest cross-section between the processes that involve tree-level production of the Higgs boson. It is also the second-largest among all the processes. Its signature of two energetic forwarded jets makes it possible to tag events and distinguish Higgs decays that typically have large backgrounds, opening access to the challenging measurements of the Higgs couplings. The angular correlations between the jets make the VBF mode acquire a particular sensitivity to the charge-parity properties of the Higgs boson [23].

Higgs production in VBF is considered as a pure electroweak process at Leading Order (LO), as illustrated in Fig. 1.6. However, at the collider, the QCD corrections are typically considerable. Through the past years, considerable enhancements have been achieved in the phenomenological interpretation of the Higgs VBF production mode. Computations for the total cross-section and differential distributions to Higgs produced via VBF, including NLO QCD and EW corrections, were presented by [24, 25] and are available in flexible Parton-level MC generators. Moreover, Partial

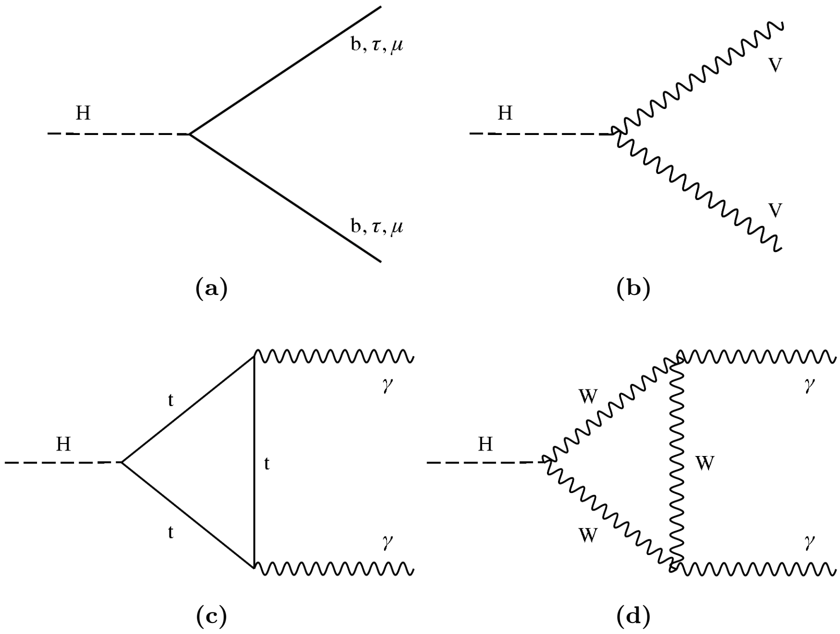


Figure 1.5: The Feynman diagrams that represent the Higgs boson decays [22]. (a), (b) are Higgs decay vertices at leading order, while (c), (d) are Higgs decays to photon pairs via loop diagrams.

NNLO QCD corrections have been presented by [26, 27], which reduces the uncertainties on the inclusive VBF XS from $5-10\%$ to $1-2\%$, making VBF the theoretically most accurate production mode at hadron colliders [10].

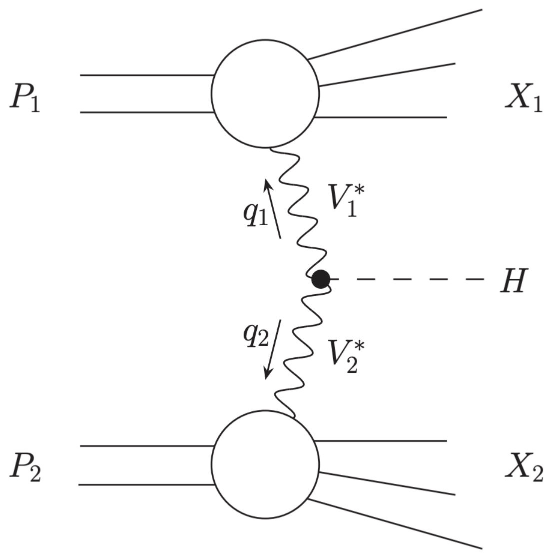


Figure 1.6: Sketch of the Higgs VBF production description through the structure function approach [27]

The most recent developments done on the VBF theory formalism come from applying the so-called structure-function approach, which is based on the description of the VBF process as a double deep-inelastic scattering process (DIS), see Fig. 1.6, in which the Higgs boson formed from the fusion of two virtual vector bosons independently emitted from the initial hadronic states. This approximation is built on the absence (or smallness) of QCD interference between the two inclusive final states X_1 and X_2. The total cross-section is given as a product of the matrix element $\mathcal{M}_{\mu\nu}$

for VBF ($V_1^\mu V_2^\nu \to H$)and the DIS hadronic tensor $W_{\mu\nu}$, giving a VBF differential cross-section of the form [25, 27],

$$d\sigma = \frac{1}{2S} 2G_F^2 M_{V1}^2 M_{V2}^2 \frac{1}{\left(Q_1^2 + M_{V1}^2\right)^2} \frac{1}{\left(Q_2^2 + M_{V2}^2\right)^2} W_{\mu\nu}(x_1, Q_1^2) \mathcal{M}^{\mu\rho} \mathcal{M}^{*\nu\sigma} W_{\rho\sigma}(x_2, Q_2^2) \times \\ \frac{d^3 P_{X1}}{(2\pi)^3 2E_{X1}} \frac{d^3 P_{X2}}{(2\pi)^3 2E_{X2}} d_{s1} d_{s2} \frac{d^3 P_H}{(2\pi)^3 2E_H} (2\pi)^4 \delta^4(P_1 + P_2 - P_{X1} - P_{X2} - P_H) \tag{1.23}$$

where S is the squared centre of mass-energy, G_F is Fermi's constant, M_{Vi} stands for the vector boson mass, $Q_1^2 = -q_i^2$ and $x_i = Q_i^2/(2P_i \cdot q_i)$ are the usual DIS variables. The theoretical discussion about the computation of the total cross-section exceeds the scope of this analysis and is available in Ref. [25, 26]

1.5 The experimental highlights of the Higgs boson at LHC

The LHC Run 1 was highlighted by discovering the Higgs through the $H \to 4l$ and $H \to \gamma\gamma$ decay channels and detecting its mass, the missing parameter in the SM ($\sim$ 125 GeV). The decay to vector bosons was subsequently observed with a significance of at least 5 standard deviations (σ) both in $H \to \gamma\gamma$ [28, 29], $H \to ZZ \to 4l$ [30, 31] and $H \to WW \to 2l2\nu$ [32, 33] channels by the ATLAS and CMS collaborations.

Additionally, the spin and parity of the Higgs boson were tested [34, 35]. The observations implied that the Higgs boson has spin-zero, disfavoring the spin-2 and pseudoscalar hypothesis, consistent with the pure scalar hypothesis, $J^P = 0^+$, as predicted by the SM. The decay into photon pairs excludes the spin 1 hypothesis (Landau-Yang selection rules).

Due to the larger significances of the observations of the decays to vector bosons, the consistency of the experimental data with the SSB in the electroweak sector was one of the clear answers provided by the LHC Run 1 data, while the consistency of Yukawa couplings was yet to be probed with similar accuracy. None of the fermionic decays had reached the 5σ threshold in Run 1.

With the data collected in Run 2, these decay modes lead to the most stringent measurements of the Higgs boson properties and the first Higgs differential cross-section studies in the ATLAS and CMS collaborations [36–40], in addition, it provided access to the main fermionic couplings.

LHC Run 1 Higgs results

Higgs results from the LHC Run 1 are summarized in two combination papers using data from both ATLAS and CMS experiments [41, 42]

A combined mass measurement was performed using the full Run 1 luminosity of about 5 fb^{-1} at $\sqrt{s}$ = 7 TeV and about 20 fb^{-1} at $\sqrt{s}$ = 8 TeV for both the ATLAS and CMS. Thanks to the good reconstructed mass resolution, the two decay channels $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ are used in the measurements. The resulting combined mass is,

$$m_H = 125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{syst.})\ \text{GeV}$$

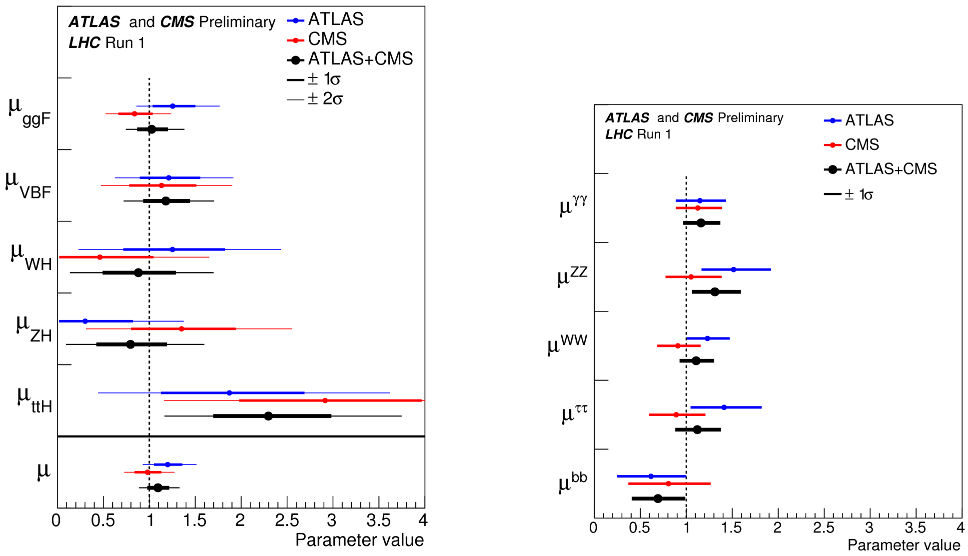


Figure 1.7: The combination of ATLAS and CMS measurements for Best-fit results for the production signal strengths (left) and the branching ratio signal (right). The error bars refer to the 1σ and 2σ intervals. [41]

The couplings to SM fermions and bosons were subsequently extracted, assuming a Higgs mass of 125.09 GeV. Several parameterizations were used for the combined fits. The signal strength modifier μ, describe the agreement between the SM prediction and the relative measurement, gives a best-fit value of

$$\mu = 1.09^{+0.11}_{-0.10} = 1.09^{+0.07}_{-0.07}(stat.)^{+0.04}_{-0.04}(exp.)^{+0.07}_{-0.07}(theory)$$

Analogous measurements were performed independently for each production signal strength and each branching ratio signal strength. The fit results are shown in Fig. 1.7.

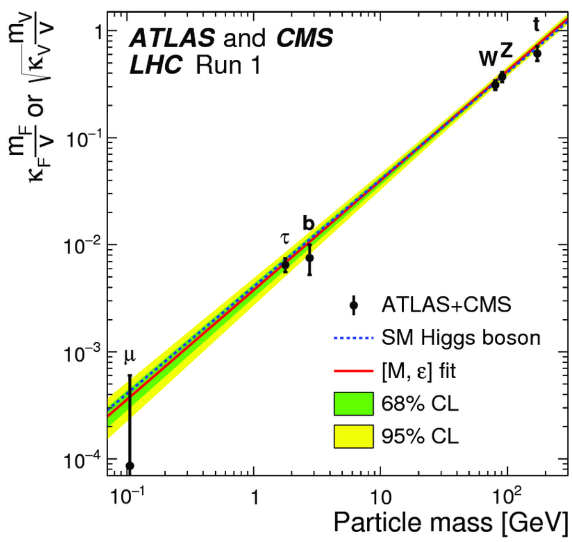


Figure 1.8: Best fit result for the coupling modifiers combining ATLAS and CMS measurement for each SM particle as a function of the particle mass; the expectation from the SM is represented by the dashed line. [41]

Fits with only the SM vertices modifiers can also be performed. In this case, the SM assumptions about the relative contribution of each coupling to the amplitudes are essential. The coupling to each SM particle is tested individually, assuming the loop amplitudes' relative consistency with the SM. Fig. 1.8 reports the performed fit that assesses the consistency of the data with the SM vertices as a function of the vacuum expectation value ν. The results show a good consistency with the expectations for both the couplings to fermions and vector bosons with relatively better precision.

LHC Run 2 Higgs results

The larger amount of data collected during Run 2, corresponding to an integrated luminosity of 137 fb^{-1}, allowed the observation of some of the decay and production modes, which were missing after Run 1. In particular, observing the Higgs to fermion couplings was crucial for verifying the Higgs sector in the SM.

The $H \to \tau\tau$ decay observation was reached after the ATLAS and CMS results for Run 1 data combinations. With Run 2 using 2016 data, both CMS and ATLAS observed this decay independently [43, 44]. Such a result confirmed the coupling of the Higgs with leptons.

The first direct measurement of the coupling of the Higgs boson with the t-quarks has been measured by observing the $t\bar{t}H$ production mode, certificating the interaction between the Higgs and the up-type quarks. Both ATLAS and CMS obtained the 5σ significance over the background-only expectation during Run 2 [45, 46].

The observation of $H \to b\bar{b}$ was achieved in 2018 by CMS and ATLAS using 2016 and 2017 data [47, 48]. The most sensitive production mode for this search is VH, which exploits the vector boson's leptonic decays and allows the observation's attainment. These measurements probed the coupling of the Higgs boson to the third generation fermion.

The evidence of the Yukawa couplings of the Higgs boson with the second generation of fermions is one of the most important achievements for LHC Run 2, which has been performed through the decay channel $H \to \mu^+\mu^-$. An observed (expected) signal significance of 2.0σ (1.7σ) has been obtained from the ATLAS experiment using Run 2 data [49], followed by 3.0σ $(2.4)\sigma$ has been achieved by the CMS [50]; thus, it constitutes the first evidence of the Higgs boson decay to muons.

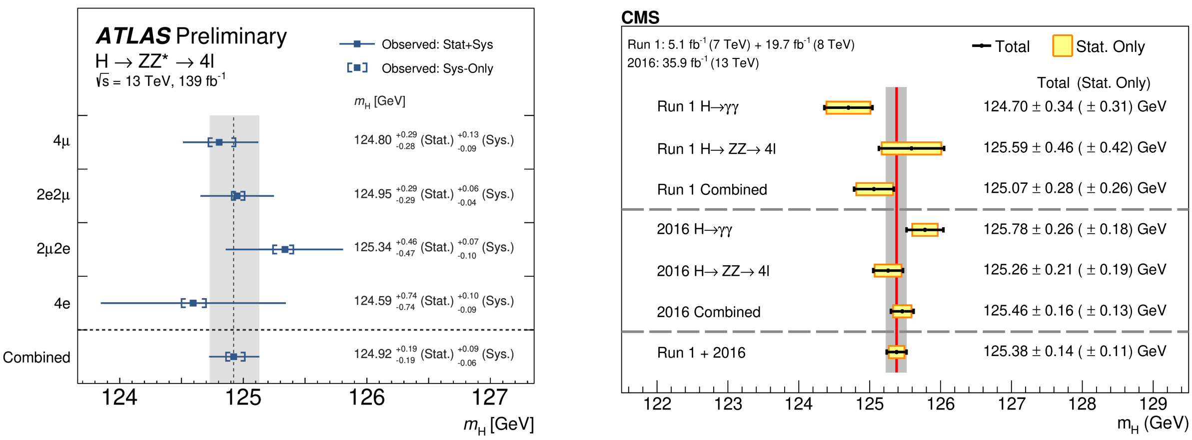


Figure 1.9: A summary of the measured Higgs boson mass obtained by the ATLAS in the decay channel $H \to ZZ \to 4l$ (left) and obtained by the CMS in the decay channels $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ (right). The results by ATLAS using Run 2 data and for CMS use Run 1 + 2016 combined measurement. [38, 51]

Several other important measurements about the Higgs, thanks to the huge amount of data collected in Run 2. In particular, the Higgs mass has been measured with a precision reaching 0.1% by CMS and ATLAS, as shown in Fig. 1.9. The most precise measurement of the Higgs boson mass up to date done by the CMS in $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ decay channels using Run 1 + 2016 combined data is

$$m_H{=}125.38 \pm 0.11(\text{stat.}) \pm 0.08(\text{syst.})\ \text{GeV}$$

The total width of the Higgs boson is also a challenging measurement at the LHC. This measurement has been performed by comparing the on-shell and off-shell production rates in the ZZ decay mode [52]. The Run 2 CMS results improved on the upper limit set in Run 1, and a lower limit is also set for the first time. The Higgs width has been measured to be $3.2^{+2.8}_{-2.2}$ MeV, in agreement with the expectation. The observed width is presented in Figure 1.6 with the 68% and 95% confidence levels.

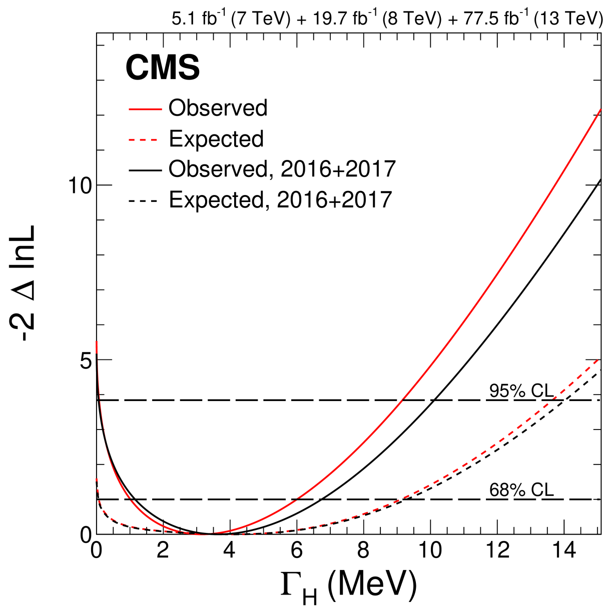


Figure 1.10: The likelihood scans of the Higgs total width, observed (solid) and expected (dashed). Results are shown using the data collected in 2016 and 2017 (black) and combined with Run 1 (red). [52]

Previous VBF H $\to ZZ \to 4l$ measurements

The VBF channel is a direct probe of the coupling between vector bosons and the Higgs boson and, hence, directly probes the electroweak sector of the SM, considering the decay mode $H \to ZZ \to 4l$, it is a clean channel with very distinctive kinematics which allow precision measurements of Higgs properties. In addition, it is followed by two energetic jets separated by a η gap, which can be used as a tagging input.

The latest measurement for the signal strength modifiers for the different main production mechanisms of the Higgs boson recently performed by the CMS using Run 2 data has been shown in

Fig. 1.11, including the signal strength modifier for the VBF production mode. From the measurements, the best fit for the expected signal strength modifier for the VBF production mode at m_H= 125.38 GeV is $\mu = 1.00^{+0.53}_{-0.44}(stat)^{+0.18}_{-0.12}(syst)$, while the observed is $\mu = 0.48^{+0.37}_{-0.46}(stat)^{+0.14}_{-0.10}(syst)$.

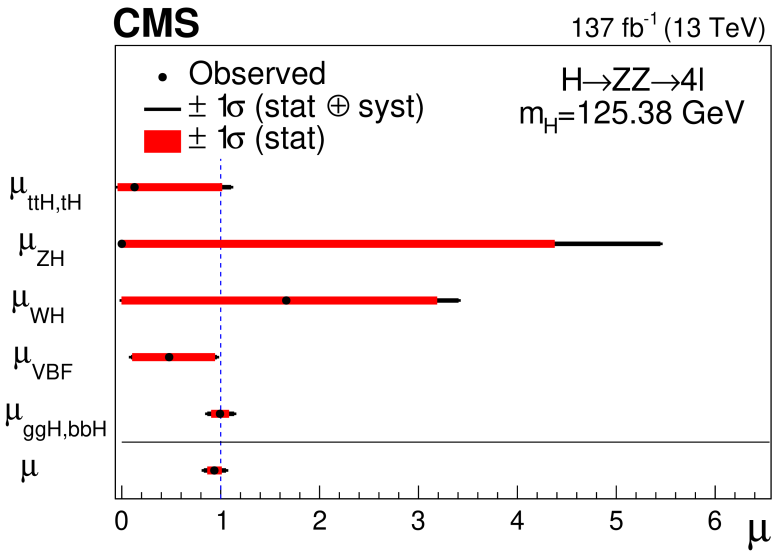


Figure 1.11: The signal strength modifiers for the different main production mechanisms of the Higgs boson, compared to the SM prediction represented by a vertical dashed line. The red lines show the statistical uncertainties corresponding to the 1σ confidence intervals, while the black lines show the σ confidence intervals, including statistical and systematic sources. [37]

Subsequently, a great effort has been devoted to the analysis of the full CMS Run 2 dataset to improve the measurements for the VBF Higgs production mode, which is the main topic of this thesis. Several improvements were added to the different aspects of the analysis and new techniques used, as will be discussed in detail in chapters 5, 6, 7, 8.

Chapter 2

The CMS detector at LHC

The analysis performed in this thesis is done using proton-proton collision data collected by the CMS detector during the LHC Run 2. This chapter provides an overview of the LHC accelerator and the CMS detector. Section 2.1 includes the design, the parameters and the operations of the different phases of the LHC. The CMS subdetector structure is discussed in Section 2.2, and the CMS trigger system and the relative experimental challenges are covered in Section 2.3. Then the algorithms used to reconstruct and identify the physics objects used in this analysis are described in Section 2.4. Finally, the simulation of the interaction of the particles in the CMS detector, introduced in Section 2.5.

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is the largest and most powerful particle accelerator ever constructed. It is located underground at an average depth of 100 m, straddling the border between Switzerland and France, at the CERN laboratories [53]. The LHC was conceived to investigate the nature of the SSB by pursuing the search for the Higgs boson and scan the entire accessible space, searching for new phenomena. Furthermore, a physics program of heavy ions collisions is also carried out, focusing on studying the collective behaviour of quarks and gluons in the form of plasma.

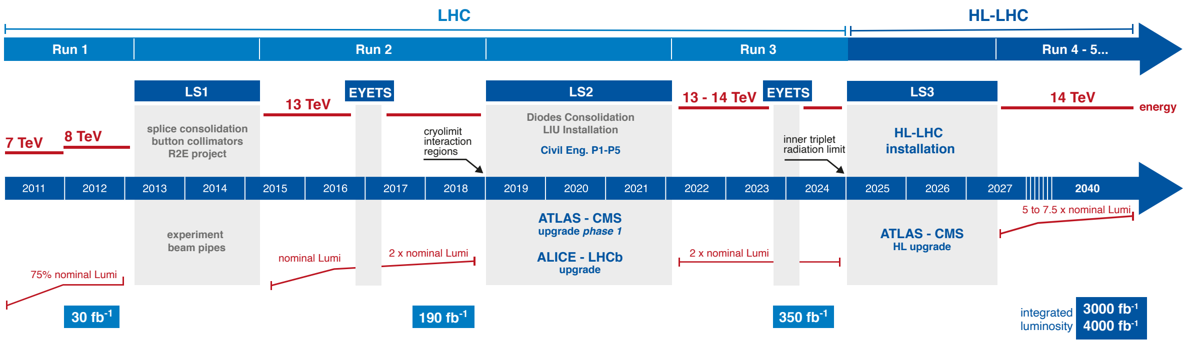


Figure 2.1: The baseline plan for the LHC and the HL-LHC for the next decade and beyond operation showing the evolution of the beam energy and integrated luminosity. "EYETS" indicates an extended year-end technical stop. After the long shutdown 3 (2025-2027), the LHC will be in the high-luminosity (HL-LHC) configuration. [54]

The LHC collides the particle beams with high center-of-mass energy enough to produce new heavy particles and deliver several collisions that allow very rare processes to be observed. These requirements drive the machine design and the beam structure, which will be described in this section. Its physics operations started in 2009 and will last until 2024 in the current configuration; then, the machine will undergo a significant upgrade towards the High Luminosity LHC phase (HL-LHC). The LHC timeline with the evolution of the beam energy and integrated luminosity is shown in Fig. 2.1.

2.1.1 Particle accelerator complex at CERN

Before accessing the LHC, the proton beams are accelerated in a complex system of smaller machines; at each step of the acceleration chain, the beams are boosted to higher energy exploiting the previous machine as an injector. The CERN accelerator complex is represented in Fig. 2.2.

At the beginning of the chain, protons are produced by ionizing hydrogen gas using a large electric field; the electrons are torn off the hydrogen atoms yielding protons. Those protons are then accelerated in the Linac 2 linear accelerator, which brings the proton beam to an energy of about 50 MeV. They are then injected into the first circular collider, the Proton Synchrotron Booster (PSB), which accelerates the beam up 1.4 GeV and passes it to the Proton Synchrotron (PS) for three to four cycles that bring the beam to about 25 GeV. Afterwards, the proton beam is injected into the last pre-acceleration step before the LHC, the Super Proton Synchrotron (SPS), which accelerates the protons to 450 GeV in about twelve cycles.

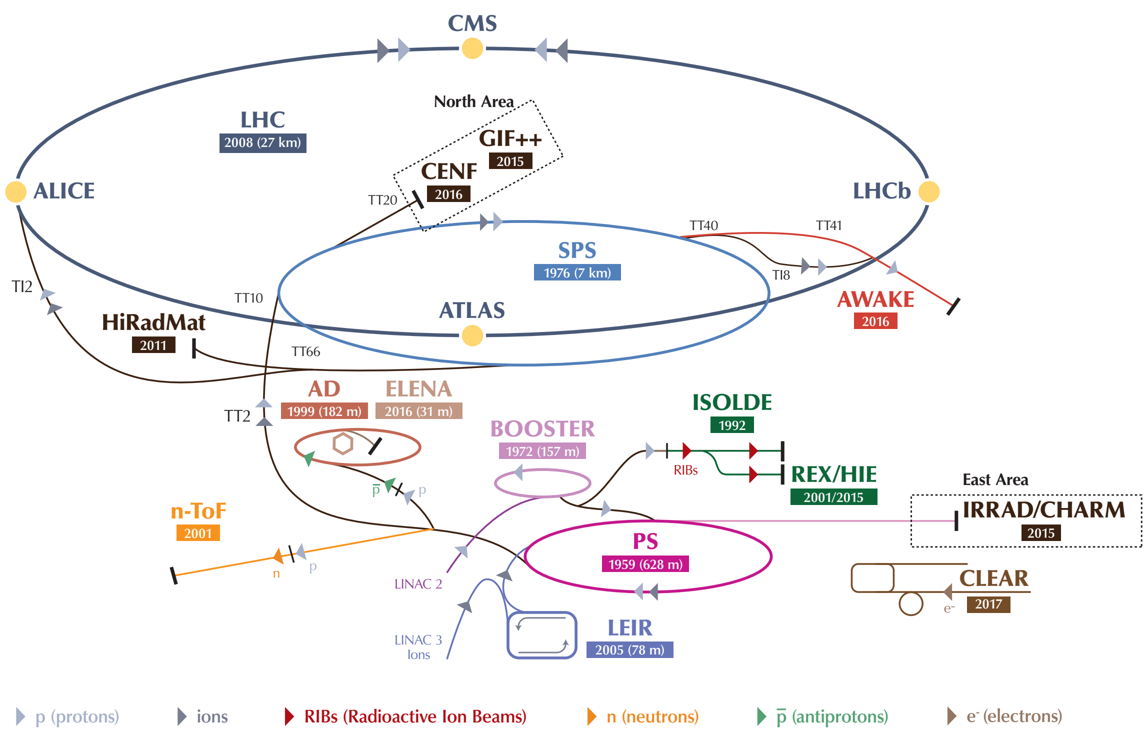


Figure 2.2: Scheme of the CERN accelerator complex [55]

The beam is eventually split into two adjacent parallel beamlines and injected into the LHC tunnel. They keep travelling in opposite directions before colliding in the designated interaction points once they have reached the maximum energy. The design center-of-mass energy for proton-proton collisions is $\sqrt{s} = 14$ TeV; nevertheless, the target energy of the LHC acceleration stage changed over the research phases, as detailed in Sec. 2.1.4.

LHC uses powerful superconducting magnets to bend particle motions effectively into a circular trajectory and focus it on the four LHC collision points. The magnetic field of 1232 dipoles operating at 8.33 T ensures the beams are kept on their circular path, while 392 quadrupole magnets keep the particles confined in narrow beams. The magnets are retained in a superconducting state using a cryogenic system that uses superfluid helium at 1.9 K temperature. Other multipole magnets are used to guarantee that bunches persist focused even with the electromagnetic repulsion of the charged particles within the beam.

2.1.2 Experiments

The beam crossing at the LHC occurs in multiple regions to collect data simultaneously in several points. There are four main experiments, with detectors installed at the collision points, as indicated in Fig. 2.2.

ATLAS [56] and CMS [57] are two general-purpose experiments with similar physics programs and different designs in order to cross-check any possible discovery. Their detectors were optimally designed to detect particles produced by Higgs boson decays, perform precision measurement of the SM, and search for any indication for a new physics. Because their searches are focused on rare processes, a large instantaneous luminosity is required at their collision points.

LHCb [58] is an experiment almost exclusively designed to study the heavy flavor quarks physics, with the principal goal of searching for evidence of new physics in CP violation in electroweak interactions.

ALICE [59] is a low luminosity experiment, collecting heavy-ion collisions data to study the quark-gluon plasma physics and the sector of the strong interaction of the SM.

In addition, there are small experiments located near the main detectors with very functional research: MoEDAL, TOTEM, LHCf, and FASER.

2.1.3 Parameters

For a specific physics process, the total number of instances where that process is reproduced ("events") N_{evt} is derived from the luminosity delivered by the machine L, and the cross-section of the physics process σ as:

$$N_{evt} = \sigma \int L \, dt \tag{2.1}$$

where L is the instantaneous luminosity representing a measure of the number of collisions occurring per second, and the integral goes over the activity time of the experiment. The luminosity is considered a key parameter of a machine; the bigger it is, allows processes with low probability to be produced. However, handling a large instantaneous luminosity is an experimental challenge for the data-taking, as debated in Sec. 2.1.4.

Assuming that the two beams are identical, have round transverse sections and are highly collimated, L can be written as:

$$L = \frac{N_p^2 n_b f_{LHC} \gamma_r}{4\pi\epsilon_n\beta^*} F, \tag{2.2}$$

where F is a geometric factor accounting for the luminosity reduction factor due to the crossing angle at the interaction point and is calculated as:

$$F = \left[1 + \left(\frac{\theta_c \sigma_z}{2\sigma^*}\right)^2\right]^{-1/2} \tag{2.3}$$

Table. 2.1 summarises the LHC designed parameters and their meaning.

The average number of simultaneous interactions that occur per bunch crossing, i.e. the pileup, is given by

$$< PU >= \frac{L \cdot \sigma_{pp}^{inel}}{n_b \cdot f_{LHC}} \tag{2.4}$$

where σ_{pp}^{inel} is the inelastic pp cross-section. From eq. 2.4 the pileup can be kept under control by managing the nominal number of bunches and the instantaneous luminosity. The average pileup in nominal conditions is about 22. However, the nominal values were exceeded during ordinary 2017 operations, as discussed in Sec. 2.1.4.

Parameter	Meaning	Nominal value
$\sqrt{s}$	Center-of-mass energy	14 TeV
Δt	Bunch separation	25 ns
n_b	Number of bunches	2808
N_p	Number of protons per bunches	1.15 $\times 10^{11}$
f_{LHC}	Revolution frequency	11245 Hz
σ^*	Transverse bunch r.m.s. at the interaction point	16.7 μm
σ_{xy}	Longitudinal bunch r.m.s.	7.55 cm
β^*	Beta function at the interaction point	0.55 m
θ_c	Crossing angle at the interaction point	285 μrad
ϵ_n	Transverse emittance	3.75 μm

Table 2.1: Nominal LHC parameters in proton-proton collisions.

2.1.4 Operations

The first run of LHC started officially in November 2009 with collisions at a center-of-mass $\sqrt{s} = 0.9$ TeV. In March 2010, the LHC began its research program with the first-ever collisions at $\sqrt{s} = 7$ TeV. Two eras of LHC physics operations are already concluded: the Run 1 and Run 2 phases. The center-of-mass energy and the instantaneous luminosity have been increased over the years towards the nominal performance level (see Fig. 2.1). The evolution of the instantaneous and integrated luminosity recorded by CMS is summarised in Fig. 2.3; the average pileup distribution per year is presented in Fig. 2.4.

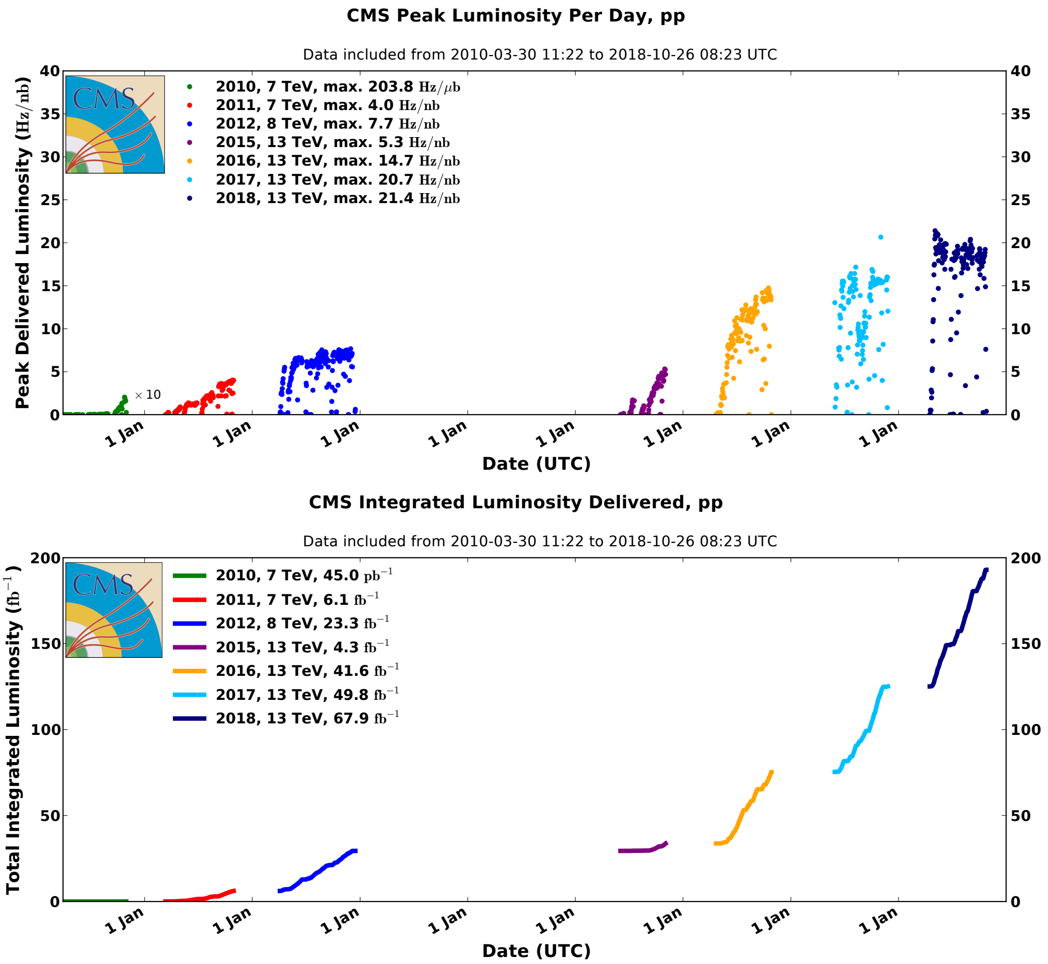


Figure 2.3: Peak instantaneous luminosity (top) and total integrated luminosity (bottom) recorded by the CMS experiment during Run 1 and Run 2 for pp collisions. [60]

At the beginning of the Run 1 data-taking period, the proton beams were accelerated to 3.5 TeV each to reach $\sqrt{s} = 7$ TeV collisions; by the end of 2011, a total integrated luminosity of 6.1 fb^{-1} was delivered. In 2012, the collisions took place at $\sqrt{s} = 8$ TeV, and a larger set of data about 23 fb^{-1} was recorded. These operations lead to the observation of the Higgs bosons, announced by ATLAS and CMS Collaborations in July 2012.

Run 1 was followed by the first Long Shutdown (LS1), which lasted about two years. During this time, the consolidation work for the accelerator was performed to raise the beam's energy to 6.5 TeV. In addition, the necessary detector upgrades have been done for the experiments to cope with the increasing luminosity.

All along the Run 2, lasted from 2015 until 2018, the collisions took place at 13 TeV. The 2015 operations correspond to a phase of commissioning of the new energy configuration. In 2016, the luminosity for proton-proton collisions increased up to 40% above the design value. In 2017 the luminosity was increased more and reached twice the design value. At the end of 2018, Run 2 terminated, and both ATLAS and CMS were able to successfully record and reconstruct the vast majority of the delivered luminosity around $\sim 140\ fb^{-1}$ of data.

The LHC is currently in a phase of maintenance and upgrades (LS2) in preparation for Run 3 when the collisions take place at the designed value $\sqrt{s} = 14$ TeV; the expected integrated luminosity to be delivered by the LHC during Run 3 amount to 300 fb^{-1}.

Run 3 will conclude the LHC Phase 1 and then the LHC and the accelerator complex will undergo a profound upgrade towards the High Luminosity LHC (HL-LHC) during the LS3, entering Phase 2. By the end of Phase 2 operations, the total integrated luminosity is expected to exceed 3000 fb^{-1}.

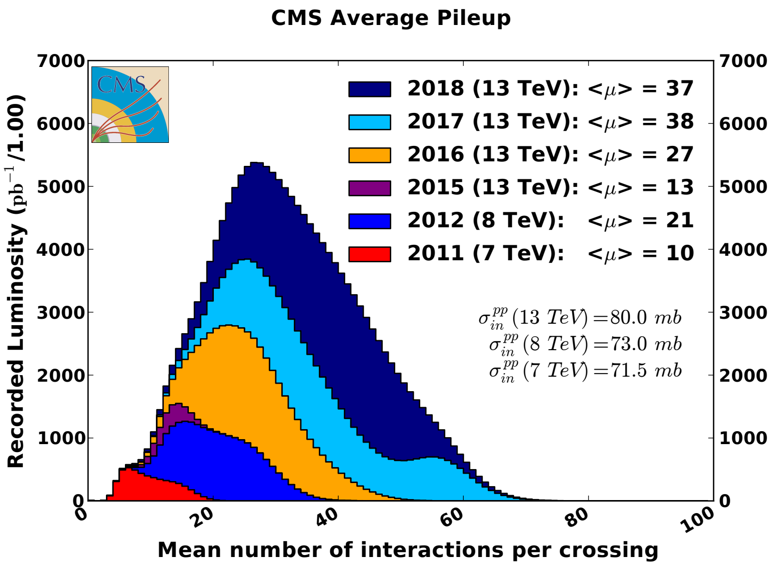


Figure 2.4: Distribution of the average pileup vertices for the different data-taking years recorded by the CMS. [60]

2.2 The Compact Muon Solenoid Detector

The CMS experiment uses a multi-purpose detector designed aiming at the best resolution for the reconstruction of the Higgs boson decay products. It comprises several subdetectors placed concentrically around the interaction point, in a cylindrical configuration, covering most of the solid angle around it; a schematic view for the CMS is shown in Fig. 2.5. The central part of the CMS apparatus is called "barrel", while the extremities covering the forward regions are called "endcaps". As its name points, the CMS detector is relatively compact, given its size and the amount of material that it is made of: being 21.6 m long and has a diameter of 14.6 m, it is smaller than the ATLAS detector by about a factor of two; nevertheless, it is almost twice as heavy, with a weight amount to 12500 tonnes.

Each subdetector is designed to perform a specific task and is dedicated to identify and reconstruct a particular type of particle coming out of the collision. **The muon system** is dedicated to identifying muons and the precise measurement of their momenta. **The electromagnetic calorimeter (ECAL)** measures the energy of particles that interact electromagnetically, primarily electrons and photons, whereas **the hadronic calorimeter (HCAL)** is designed to detect charged and uncharged hadrons; i.e., protons, neutrons, pions, and kaons. **The tracker** detects all the charged particles, either isolated or in jets, with excellent spatial resolution. Neutrinos are detected as

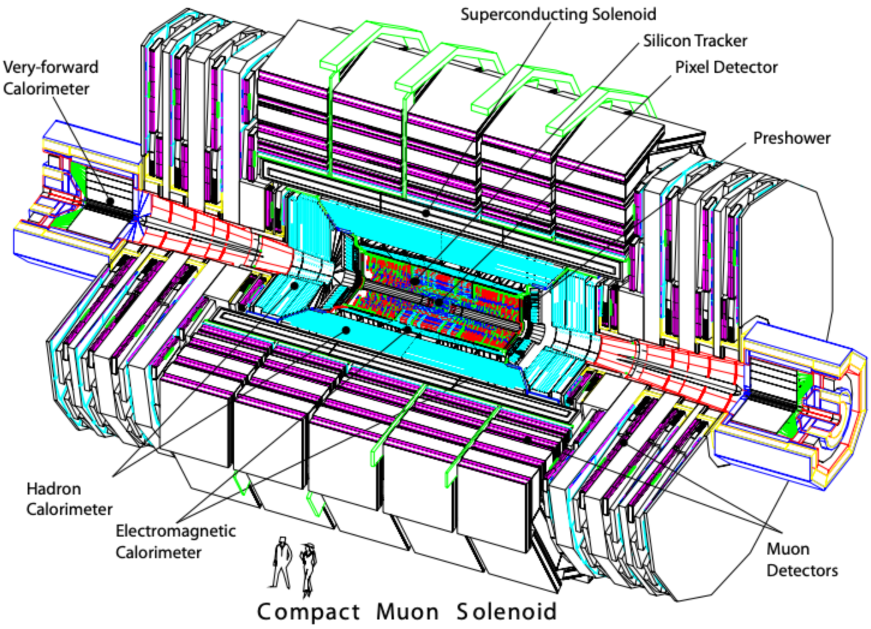


Figure 2.5: Schematic view of the CMS experiment and its subdetectors [57]

missing transverse energy. The different CMS subdetectors have been described in the following sections, starting from the tracker, which is the closest to the interaction point, moving outwards.

Coordinate system

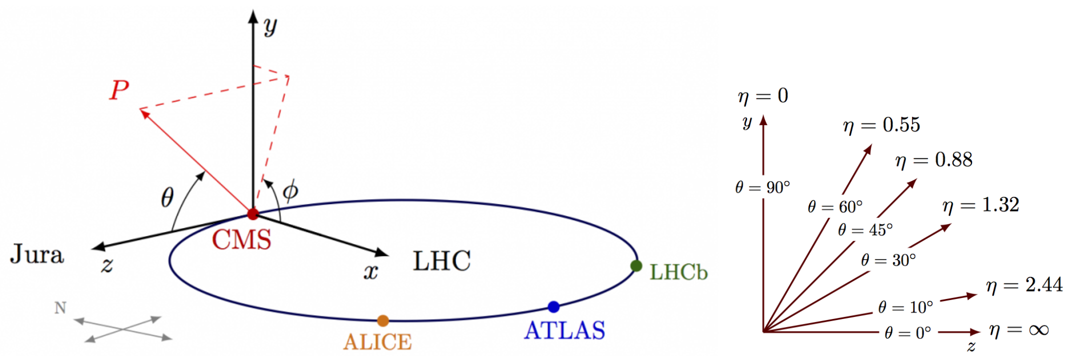


Figure 2.6: Illustration of the CMS coordinate system (left) and the relation between the pseudorapidity η and the polar angle θ (right). [61]

To describe the positions and directions of reconstructed objects, the CMS uses a coordinate system centered in the nominal interaction point. The x-axis points to the center of the LHC, the y-axis is oriented upwards perpendicularly to the LHC plane, and the z-axis is perpendicular to the transverse plane (x-y), parallel to the beam axis with a positive, counter-clockwise direction, as illustrated in Fig. 2.6 (left). The azimuthal angle ϕ is measured from the positive x-direction in the plane (x-y), and the polar angle θ is measured from the positive z-direction.

Rather than using the polar angle θ to indicate directions, the pseudorapidity relative to the z-axis is defined as:

$$\eta = -\ln \tan \frac{\theta}{2} \tag{2.5}$$

which varies from 0 for $\theta = \pi/2$ and infinity for $\theta = 0$, as represented in Fig. 2.6 (right). The detector regions with large η are referred to as the "forward regions". Pseudorapidity differences are invariant under Lorentz boosts along the z-axis. Thus, the spatial separation between two particles is represented as:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \tag{2.6}$$

The solenoid magnet

The main feature of the CMS detector is the presence of the central superconducting solenoid magnet, operating at a 4.5 K temperature, produces an axial and uniform magnetic field of 3.8 T along the z-axis. The solenoid constraints drive the whole CMS detector design; installing the magnet outside the tracker's volume and calorimeters minimizes the amount of material placed in front of the calorimeters, facilitating the link between tracks and calorimeter clusters. The density of the calorimeters material must then be very high to allow for the electromagnetic and hadronic showers to be exhausted within the volume bounded by the solenoid. The track bending guaranteed by the intense magnetic field is powerful enough to separate between the neutral and charged particles.

2.2.1 The Tracking system

The CMS tracking system [62, 63] is the innermost detector, closer to the interaction point. It is dedicated to provide a precise and efficient measurement of positions and momenta of charged particles produced from the collision, and consequently for primary and secondary vertex reconstructions. It is immersed in a magnetic field of 3.8 T provided by the CMS solenoid.

The tracker needs some specific requirements in order to achieve the best performance: It has to be light to avoid altering charged particles' trajectories; fast enough to take data every 25 ns (40 MHz); large hit redundancy to perform good pattern recognition; provide the best possible resolution, good radiation hardness, high granularity is also necessary due to the high particle multiplicity.

The silicon technology has been chosen to build the entire tracker to satisfy the mentioned requirements: a silicon pixel detector is used in the inner part, most near the interaction point, and silicon microstrip detectors are installed in the outer region. The total length of the tracker is 5.8 m, its diameter is 2.5 m. The layout of the original CMS tracker is presented in Fig. 2.7.

In March 2017, the original pixel detector, shown in Fig. 2.7, was replaced with a new device, the "phase-1" pixel detector [64], equipped with one additional barrel layer and one additional forward disk per side that features full 4-hit coverage in the tracking volume. This upgraded design places the innermost barrel layer closer to the interaction point, 2.9 cm instead of 4.4 cm. The main motivation was to address the dynamic inefficiency driven by the internal readout chip buffers at high instantaneous luminosity and trigger rates. A comparison between the geometry of the old

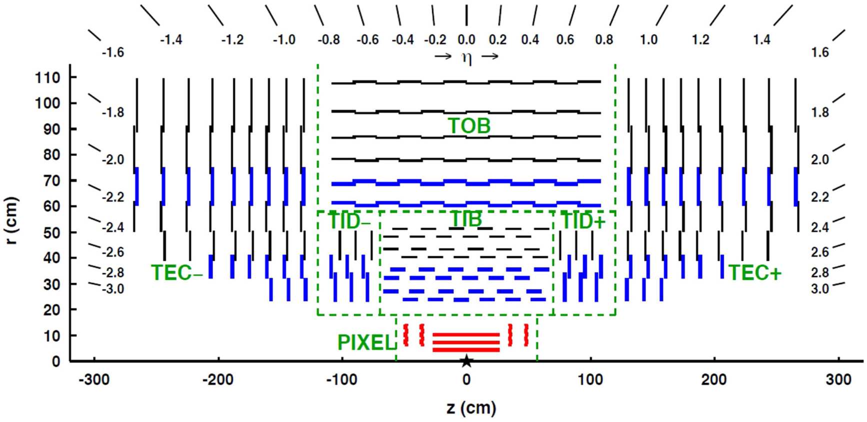


Figure 2.7: Schematic view of the CMS original tracker layout in the r-z plane. [63]

and new pixel detectors is shown in Fig. 2.8. The pixel size is unchanged: $100 \times 150\,\mu m^2$. The hit position resolution in the pixel tracker is around $10\,\mu m$ in the transverse coordinate and $20-40\,\mu m$ in the longitudinal coordinate. The current pixel detector is planned to be used through the LHC Run 3 with the original outer tracker.

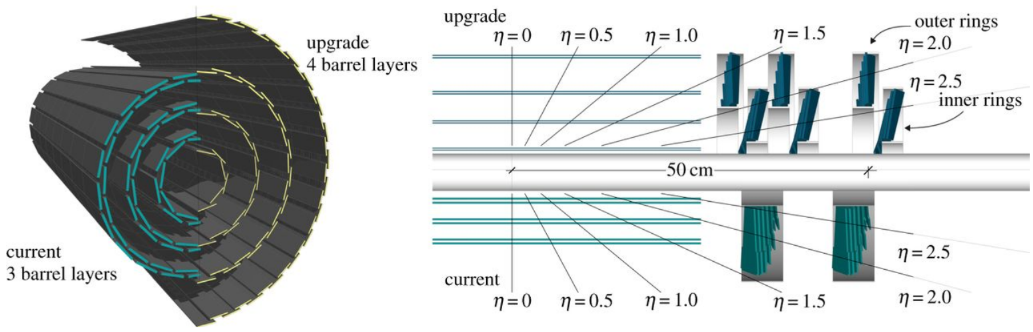


Figure 2.8: Transverse-oblique view comparing the pixel barrel layers of the initial and upgraded CMS pixel detectors (left). Conceptual layout comparing the layers and disks in the two detectors (right). [65]

Surrounding the pixel detector is the strip tracker, it consists of four subsystems: tracker inner barrel (TIB) and tracker outer barrel (TOB) with cylindrical shape placed in the barrel region, and tracker inner disk (TID) and tracker endcap (TEC) installed in the two endcap regions. The location and number of layers for each subsystem are summarized in Table 2.2. The single hit resolution in TIB (TOB) is about $13-38\,\mu m (18-47\,\mu m)$.

Subsystem	Layers	Location
TIB	4 cylindrical	$20 < r < 55$ cm
TOB	6 cylindrical	$55 < r < 116$ cm
TID	3 disks	$58 < r < 124$ cm
TEC	9 disks	$124 < r < 282$ cm

Table 2.2: The number of layers and the position information for the four components of the strip tracker.

The four regions are provided with both single-sided and double-sided microstrip modules. The double-sided modules are rotated with respect to the strip direction by a "stereo" angle of 100 μrad. They can therefore provide 3D measurements, though with $\sim$ 10 times lower resolution along the strips. The strips are oriented along the z-axis and the r-coordinate in the barrel and endcap regions, respectively. The microstrip detector design spatial resolution is about $20-50\,\mu m$ in the transverse direction and $200-500\,\mu m$ along the strips for "stereo" hits. The performance of the tracking system will be discussed in detail in chapter 3.

2.2.2 The Calorimeters

The calorimeters are located outside the tracker and inside the solenoid magnet. They are designed to measure the energy of both electromagnetic and hadronic showers. Unlike the tracking system, they are required to completely absorb the particles in the shower for optimal energy measurements: it measures the energy lost by incident particles due to interactions with the detector material, which helps in identifying the incident particle based on the depth of penetration into the calorimeters and the profile of the energy lost.

2.2.2.1 Electromagnetic calorimeter

The Electromagnetic Calorimeter (ECAL) [66] is a hermetic and homogeneous array of lead tungstate (PbWO4) scintillating crystals, whose purpose is to measure the energy of the electromagnetic showers: electrons, photons, and neutral pions precisely. Lead tungstate crystals have high density (8.28 g/cm^3), short radiation length (X_0 = 0.89 cm), small Molière radius (RM = 2.2 cm), fast scintillation decay time and high radiation tolerance. These make them a suitable choice for a compact ECAL with fine granularity. Due to different magnetic fields and radiation levels, silicon avalanche photodiodes and vacuum phototriodes are chosen as photodetectors in the barrel and endcap regions, respectively.

The ECAL consists of several subsystems: the ECAL barrel (EB), the ECAL endcap (EE) at the two sides, and a sampling ECAL preshower detector (ES) in front of EE with high granularity for discriminating neutral pions from prompt photons. The layout of the ECAL is shown in Fig. 2.9.

The EB extends within the region $|\eta| < 1.479$ and is instrumented with about 61200 trapezoidal. The two EE disks have coverage of $1.479 < |\eta| < 3.0$ and contain 7324 crystals each. In both the barrel and the endcaps, crystals are oriented with their axes tilted up to 3° with respect to the direction that points to the nominal IP to avoid acceptance gaps between them. A summary of ECAL parameters is provided in Table 2.3. A more delicate electromagnetic pre-shower detector (ES) is installed in front of the two endcaps; it improves spatial resolution in the region $1.65 < |\eta| < 2.6$. It is a sampling calorimeter with two layers of lead absorber followed by 22 mm long silicon strips to measure the shower's energy deposit and transverse profile.

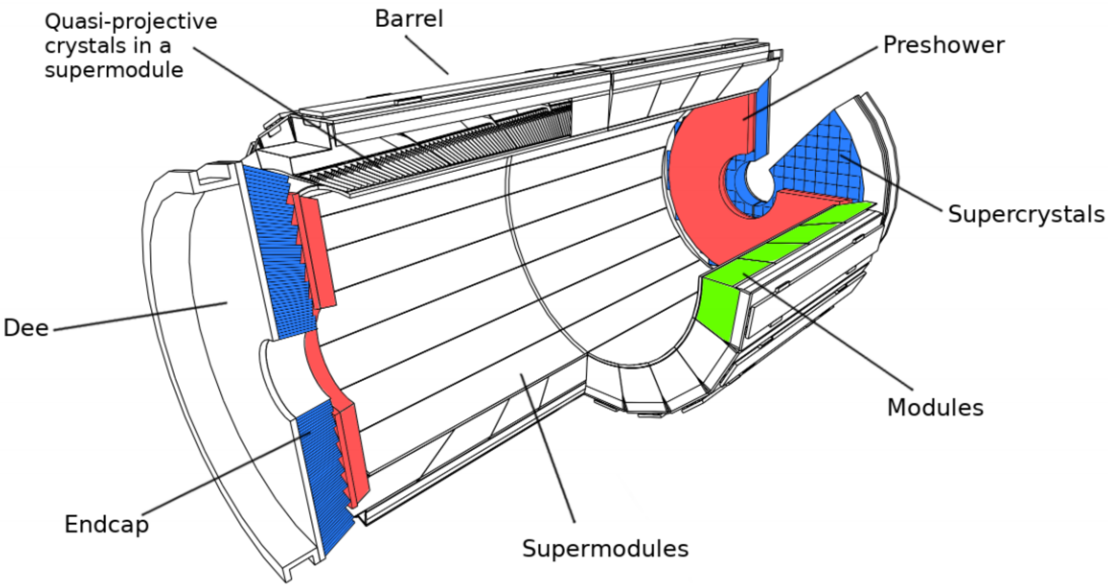


Figure 2.9: Layout of the CMS ECAL. [67]

Parameter	Barrel	Endcap
Pseudorapidity coverage	$\|\eta\| < 1.479$	$1.479 < \|\eta\| < 3.0$
ECAL envelop: $r_{inner}, r_{outer}[mm]$	1238, 1750	316, 1711
ECAL envelop: $z_{inner}, z_{outer}[mm]$	0, ±3045	±3170, ±3900
Granularity: $\Delta\eta \times \Delta\phi$	0.0174×0.0174	0.02×0.02 to 0.05×0.05
Crystal dimension $[mm^3]$	$21.8 \times 21.8 \times 230$	$28.6 \times 28.6 \times 220$
Depth in X_0	25.8	24.7
No. of crystals:	61200	14648

Table 2.3: Summary of ECAL parameters.

The ECAL energy resolution can be affected by several sources, parametrized as:

$$\frac{\sigma_E}{E} = \frac{a}{E} \oplus \frac{b}{\sqrt{E}} \oplus c \tag{2.7}$$

where a is the noise term related to the electronics and pileup, b is the stochastic term that considers mainly the fluctuations in the photon conversions, and c is a constant term that accounts for the energy scale calibration.

A typical measured PbWO4 crystal energy resolution is of the order of

$$\frac{\sigma_E}{E} = \frac{0.12}{E/GeV} \oplus \frac{2.8\%}{\sqrt{E/GeV}} \oplus 0.3\% \tag{2.8}$$

The expected performance have been almost matched during the data taking. A resolution of 1 % has been achieved for electrons with energy greater than 15 GeV.

2.2.2.2 Hadronic Calorimeter

The hadronic calorimeter (HCAL) [68] measures the energy of the hadrons, which interact mainly via the strong force. These typically traverse the ECAL volume, where they only deposit 30% of their energy and then stopped in the dense material of the HCAL, which absorbs and measures the remaining energy. The HCAL is the only CMS component capable of measuring the energy of neutral hadrons, and With a hermetic structure, it plays a vital role in reconstructing and measuring the energy of neutrinos and angular coverage in the forward region for the forward jets.

The HCAL consists of four subsystems: the barrel HCAL (HB) surrounds the ECAL and covers the central pseudorapidity region up to $|\eta| = 1.3$; the two endcap HCAL (HE) cover up to $|\eta| = 3$. The HCAL forward (HF) extends the coverage up to $|\eta| = 5$ in the forward region. An array of scintillators, the outer HCAL (HO), is located outside the magnet, uses the magnet material as an absorber to catch the tails of the hadronic shower, and avoids the misidentification of muons. A layout of the HCAL with these different subsystems is shown in Fig. 2.10.

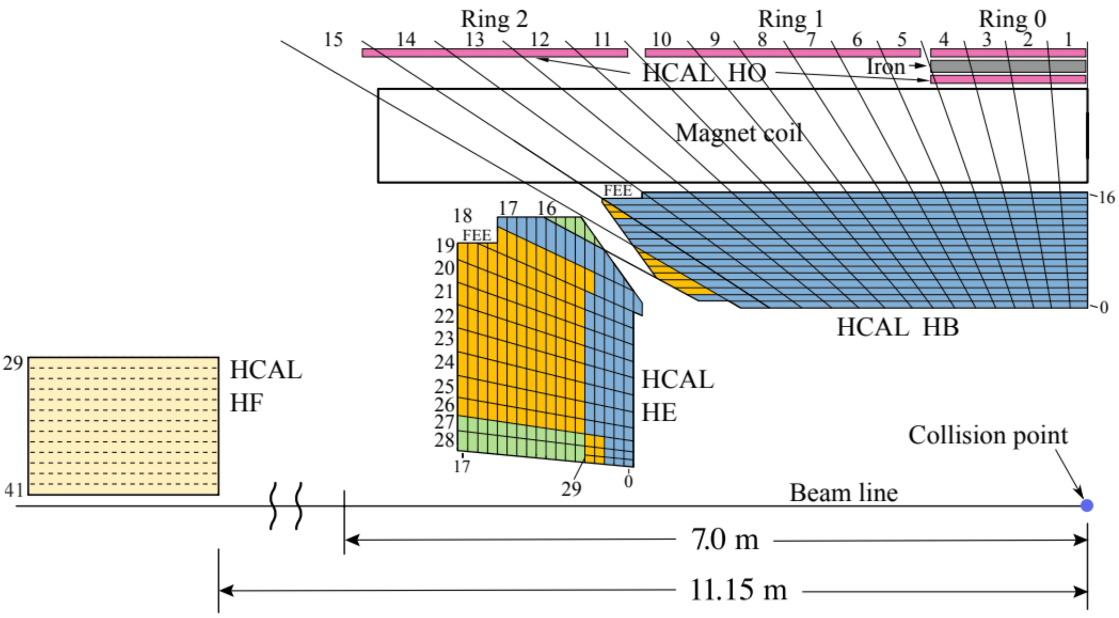


Figure 2.10: A schematic view for one-quarter of the HCAL, displaying the positions of its four main parts: the hadron barrel (HB), the hadron endcap (HE), the hadron outer (HO), and the hadron forward (HF) calorimeters. [69]

As the nuclear interaction length is much larger than the electromagnetic radiation length, the HCAL is more massive than the ECAL. The volume constraints of the magnet significantly limited the size of the HCAL; the relative size of the material ingredients was optimized for maximal shower containment. The HCAL is a sampling calorimeter consisting of alternate layers of brass absorbers and plastic fluorescent scintillators. The presence of brass as an absorber guarantees the containment of the hadronic shower, with a thickness of at least six interaction lengths. Steel interlayered with a quartz-fiber Cherenkov calorimeter is used alternatively in the forward region. The HO detectors have no dedicated absorbers, but the shower is induced by the magnet and return yoke material.

The combined energy resolution of ECAL + HCAL has been measured to be [70]

$$\frac{\sigma_E}{E} = \frac{84.7\%}{\sqrt{E}} \oplus 7.4\% \tag{2.9}$$

The HCAL photodetectors, readout system, and electronics were upgraded at the end of 2017 [71]. The HF detector initially used photomultiplier tubes to collect the Cherenkov light, which is now replaced by multi-anode tubes to reduce the rate of anomalous signals; With introducing precision timing measurements and longitudinal depth segmentation, the pattern recognition capabilities in the HCAL were improved, granting an additional handle for background rejection.

2.2.3 The Muon System

As the name and the logo of CMS reflect, the precise measurement of the muons momenta is a core specification of the CMS design. Muons are produced in many physics processes, and their clean and distinctive signature makes them a suitable tag for trigger decisions. Unlike electrons, photons, or hadrons, muons are less subject to radiative effects and can cross through the detector mostly unaffected. For this reason, the muon detector system [72] was installed in the outermost part of the CMS detector, outside the magnet volume. Three steel layers are present both in the barrel and in the endcap regions, alternated with four layers of muon chambers, as shown in Fig. 2.11.

The muon system provides information to identify muons and to measure the charge and momentum of high-p_T muons. Additionally, two tasks are achieved by the muon system due to its good time resolution: muon-triggering and bunch crossing identification. The muon chambers are instrumented with three different kinds of gaseous detector technologies; they are positioned throughout the muon system depending on the expected background rates and the uniformity of the magnetic field. They are arranged to provide broad coverage with spatial overlap to maximize reconstruction efficiency: **drift tube (DT) chambers** and **cathode strip chambers (CSC)** detect muons in the regions $|\eta| < 1.2$ and $0.9 < |\eta| < 2.4$, respectively and they are supplemented by a system of **resistive plate chambers (RPC)** covering the range $|\eta| < 1.6$.

The barrel region is characterized by a relatively low particle rate and a uniform magnetic field; accordingly, the DT chambers with long aluminum drift cells are used. The basic constituent of the DT detector is a rectangular cell of size 1.3×4.2 cm^2 and length 2.4 m, filled with a mixture of Ar (85%) and CO_2 (15%). Placing the electrodes on the top and bottom of the cell ensures a constant field and a uniform drift velocity of about 55 μm/s, whereas the cathodes are placed on the sides. The position of the muon traverses the gas is determined by measuring the drift time to the anode wire of the cell. There are 250 DT chambers in total, with each providing a spatial resolution of about 100 μm in $r - \phi$ and 150 μm in the z-direction.

The endcap regions are equipped with CSCs, where the background rates are higher than in the barrel, and the magnetic field is non-uniform. Each endcap has 4 stations of chambers perpendicular to the beam. A CSCs consists of 6 layers of anode wires and cathode strips, filled with a mixture of Ar (45%), CO_2 (50%), and CF_4 (10%), each measuring the muon position in 2 coordinates. The gas is ionized as the muon passes through the chamber, and the induced signals are combined to provide a precise measurement of the muon position. Cathode strips are positioned radially and

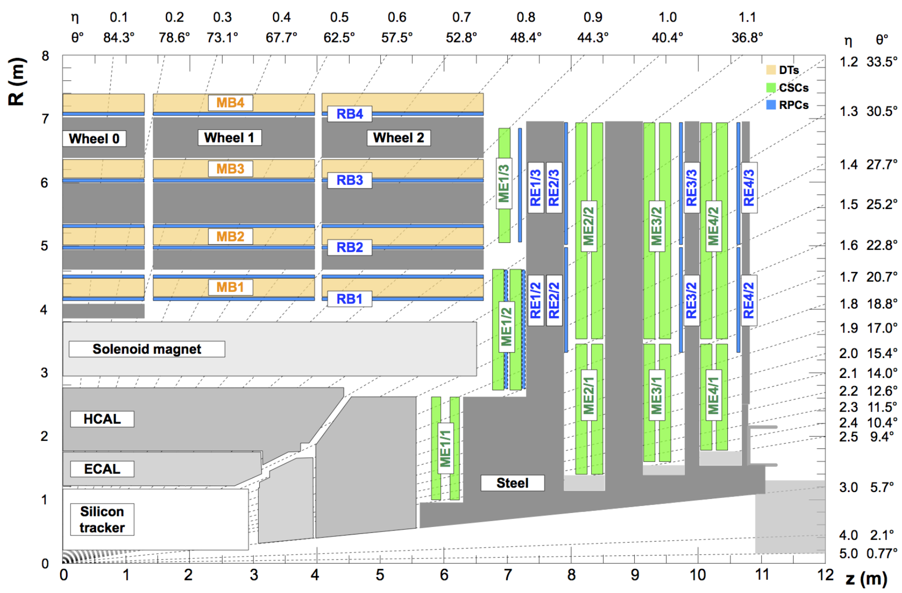


Figure 2.11: An r-z Schematic view of the CMS Muon detector, including the subsystems: DTs, CSCs, and RPCs. [73]

provide a precision measurement in the r $\times$ ϕ plane. Anode wires provide a measurement in the radial direction. The CSC stations have r $\times$ ϕ resolution of about 75-150 μm and a z resolution of about 200 μm.

Moreover, RPCs are installed in both the barrel and the endcap regions. The RPCs are double-gap chambers operated as gaseous detectors in avalanche mode. They are formed of resistive 2 mm thick layers interlaid with a mixture of gas (95.2% of $C_2H_2F_4$, 4.5% of iC_4H_{10}, and 0.3% of SF_6). The detector is operated in avalanche mode: when traversed by a muon, an electron cascade is triggered by the high electric field inside the volume and is read out using strips oriented along z in the barrel and r in the endcap regions. Although the spatial resolution is coarser with a strip pitch of $\simeq$ 1 cm and no segmentation in the orthogonal direction, these detectors are faster than the DTs and CSCs, with a perfect time resolution better than 3 ns, thus ensuring a robust bunch crossing identification and efficient triggering.

Supplementary detector chambers were introduced in the endcap region $1.6 < |\eta| < 2.4$ at the end of 2017 to increase the muon system's redundancy in these regions. The technology used is the gas electron multiplier (GEM) [74], which can accomplish the same physics goals of the RPCs in a high radiation environment. The upgrade involved the pre-installation of 10 prototype chambers; the complete installation of the 144 detectors is ongoing during LS2 [75]. Furthermore, The readout electronics of the muon chambers were also improved as part of the CMS trigger upgrade between Run 1 and Run 2.

2.3 The CMS trigger system

The average bunch crossing rate at the core of CMS (40 MHz) and the pileup make it impossible to read out and store such vast amounts of data provided by the detector. Most of the collisions result in low-energy pp interactions, which are not interesting for the physics program of CMS.

Therefore a trigger system is essential for a high-luminosity experiment such as CMS to probe the most interesting processes. The total proton-proton cross-section is more than six orders of magnitude larger than the cross-section of interesting processes. Therefore, the data needs to be reduced and selected through a trigger system, whose crucial aspect is a fast and efficient real-time selection to events that exhibit a potential physics interest while reducing the acquisition rate by a factor 10^5. The system must be adaptable to accommodate the different data conditions and robust against the LHC's high instantaneous luminosity and pileup conditions.

The trigger selection is performed based on the kinematic properties of the produced particles. It is implemented in two consecutive steps with an increasing rate reduction. The first selection is made by the Level-1 (L1) trigger [76], encoded in custom hardware processors that select up to 100 kHz of the most interesting events with an available processing time of 3.8 μs. Following this process, the High-Level Trigger (HLT) [77] elaborates a more detailed inspection of the events on a commodity computing processor farm, reducing the rate to 1 kHz in $\sim$ 200 ms. More details about the trigger system are provided in the following.

2.3.1 The Level-1 Trigger

The input of the Level-1 trigger is the raw data from the calorimeters and the muon system. It has only 3.8 μs to make a decision, and hence the full reconstruction of the physics objects cannot be applied at this stage. The L1 trigger reduces the event rate from an input of 40 MHz to an output of about 100 kHz through. Each processing element sends its results to the next element at every bunch crossing and receives a new event to analyze. The detectors must be able to resolve in time two events, and all the signals from a single event must be synchronized. During this process, the whole detector data are stored in pipeline memories with limited size ($\sim$ 160 bunch crossings, or $\sim$ 4 μs latency).

The original L1 trigger system was developed to handle the designed LHC operations; given the instantaneous luminosities predicted for Run 2, the L1 trigger could not guarantee an adequate physics performance. A major upgrade of the L1 trigger [79, 80] was installed and commissioned between Run 1 and Run 2. To maintain the same sensitivity of Run 1 at the electroweak scale, the electronic boards were replaced by advanced mezzanine cards (AMC) equipped with powerful field-programmable gate arrays (FPGA). The new design allows for more flexibility and is more suitable to trigger complex objects. The communication among subsystems is based on MicroTCA (μTCA) technology to benefit from additional flexibility and higher bandwidth. Transferring data between the modules and the central CMS data acquisition system is performed using multi-Gb/s serial optical links to allow a faster global readout.

The schematic of the upgraded L1 trigger system for Run 2 is described in Fig. 2.12. The calorimeter trigger collects the energy deposits in the HCAL and ECAL subdetectors, while the information of the hits in the DT, RPC, and CSC subdetectors is processed in the muon trigger. In the calorimeter trigger, these objects can be electrons, photons, jets, hadronic τ, or missing transverse energy. Similarly, the hits in the muon subdetectors are used to reconstruct the muon tracks. The output of the two subsystems is received by the micro global trigger (μGT), which combines the

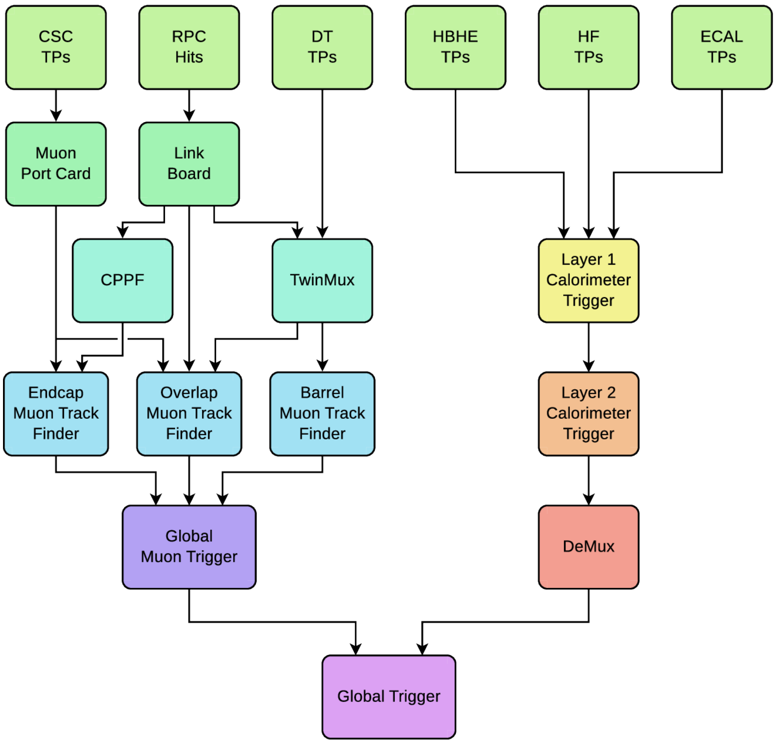


Figure 2.12: Architecture of the CMS L1 trigger system. [78]

information to make an event accept or reject decision. This selection is made on a kinematic basis: according to the candidate's energy, position, and quality criteria.

2.3.2 The High Level Trigger

The HLT further reduces the event rate of the Level-1 trigger from ∼100 kHz to ∼1 kHz, a bandwidth compatible with the data acquisition capability. The HLT is implemented by a software computing farm instrumented with more than 30000 CPUs. It takes the full detector readout as input, including tracking information, and carries out a more sophisticated reconstruction than the L1 trigger. The most stringent constraint on the HLT is the latency since the limited number of CPUs requires a maximum processing time of ∼320 ms per event. Hence, the HLT runs an online object reconstruction similar to those used for offline reconstruction, based on simplified clustering and tracking algorithms.

Reconstructed physics objects, such as leptons, photons, jets, etc., are used in the HLT. A sequence of reconstruction algorithms and selection filters for one or more physics objects is called the HLT paths. The path is characterized by the Level-1 "seed", i.e., the requirements passed by the event at Level-1, and the HLT requirements. An "HLT Menu" represents the set of trigger paths that, if enabled, contribute to the final decision that determines whether to reject or store an event. A single trigger path can require the presence of one or more physics objects of a particular type passing specific kinematic thresholds. The HLT was constantly updated to cope with the ever-changing LHC and detector operation conditions. The event rate of each trigger path should be maintained within the allowed limits given the expected instantaneous luminosity. Trigger paths with low thresholds can be kept in the HLT Menu with a "prescale" factor applied.

2.4 Physics Objects reconstruction

Each particle coming from the pp collisions at the CMS leaves a distinct signature in the subdetectors, described in Fig. 2.13. Muons lose a small fraction of energy in the innermost part of the detector, and they are detected in the tracker and the muon chambers. Both electrons and photons are absorbed within the ECAL, but only the former trajectory appears in the tracker. Hadrons traverse the ECAL, losing a small fraction of its energy and depositing most of its energy in the HCAL. The trajectory of their charged components is detected in the tracker. Neutrinos escape CMS undetected; their presence is nonetheless confirmed through the event's energy imbalance in the transverse plane. The raw information from the subdetector systems is combined to reconstruct and identify the relevant physics objects used in this analysis with maximal efficiency; the algorithms used for this purpose are briefly described in this section.

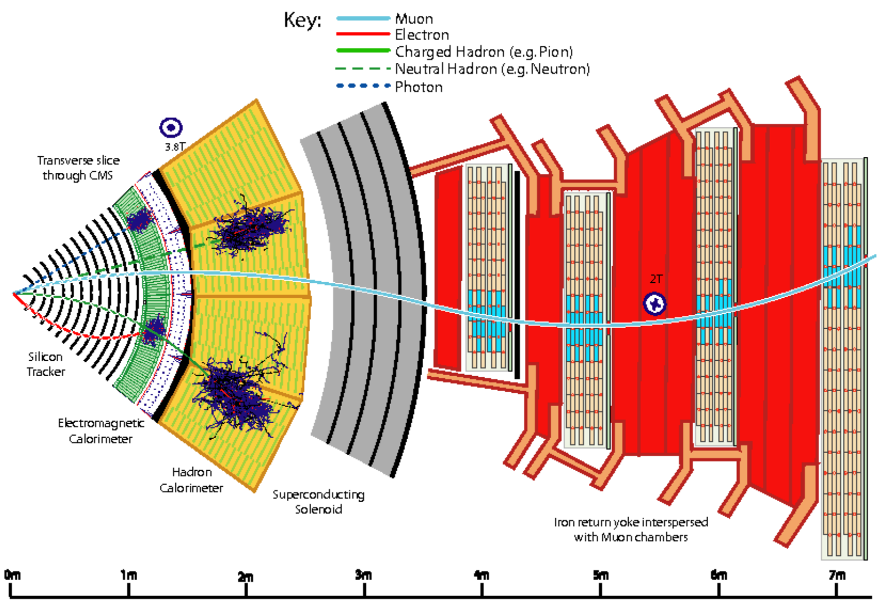


Figure 2.13: Transverse slice of the CMS detector, showing the experimental signature of the different final-state particles. [81]

2.4.1 Particle Flow basic elements

CMS uses the particle flow algorithm [81] to correlate all single detector measurements, identify the final state particles and measure their properties. The algorithm starts with the low-level information (hits in the silicon tracker and energy deposits in the calorimeters) to build basic elements (tracks and clusters) used as inputs to reconstruct the physics objects.

- **The tracks** of charged particles are reconstructed from the hits in the silicon tracker layers based on the combinatorial Kalman filter (CKF) [82] technique. The reconstruction follows an iterative approach to achieve high efficiency and a lower fake rate. The first set of iterations looks for isolated tracks that are easier to find, such as prompt tracks with relatively high p_T. Hits associated with reconstructed tracks are not considered at later stages. This procedure reduces the combinatorial complexity so that the more difficult kinematic regions tracks can be reconstructed. Typically, the track reconstruction process undergoes several iteration steps (12 iterations), enabling the algorithm to reconstruct tracks up to 0.1 GeV.

- **The clusters** are formed from the energy deposits in the calorimeters. In order to to achieve high efficiency even at low p_T and to disentangle overlapping showers, the algorithm runs separately in the preshower, the ECAL, and the HCAL subdetectors. It is based on an iterative clustering technique that follows the typical lateral shower profile. The clustering process starts with defining a cluster seed, a calorimeter cell with an energy deposit higher than the threshold and higher than the energy deposit of neighbouring cells. Then the neighbouring cells are added to the cluster until reaching a cell without any signal or have energy deposits lower than the threshold "noise level". In practice, less than five iterations are needed for the algorithm to converge.

Tracks and clusters are associated together to form PF blocks, used as inputs to reconstruct the individual physics objects. The muon reconstruction is done first by linking the PF tracks to the tracks in the muon chambers. Then, electrons are built by combining tracks to clusters with a dedicated procedure that considers the effects of bremsstrahlung radiation. The remaining clusters with no assigned track are identified as photons or neutral hadrons; the former is associated with electromagnetic deposits, and the latter with hadronic deposits. The rest of the clusters associated with a track are identified as charged hadrons. The objects reconstructed (electrons, muons, photons, charged hadrons and neutral hadrons) are combined to build more complex ones, such as jets and hadronically decaying τ leptons, as well as to estimate the missing transverse energy of the event, which indicates the presence of neutrinos.

2.4.2 Muons

Muons have very clean signatures in the muon chambers and high identification efficiency; hence they are the first particles identified in CMS. The muon tracks are reconstructed independently in the silicon tracker and the muon chambers [73]. The reconstruction of the former follows the PF procedure mentioned above and is referred to as **"tracker tracks"**. The latter is PF-independent and relies totally on the information from the muon systems indicated as **"standalone-muon tracks"**. In this case, the hits' positions within the DT, CSC and RPC detectors are associated and fitted to form track segments, which give an initial estimation of the muon's direction and momentum and then the track segments are linked to reconstruct the muon track.

Muon tracks are reconstructed with an iterative approach, running a sequence of specific tracking algorithms using inputs: the tracker tracks and muon track segments using a Kalman filter (KF) technique [82]. The global muon reconstruction in CMS proceeds by either matching tracker muons to standalone-muons (inside-out) or inversely (outside-in), as follows:

- **Tracker muon** tracks, built inside-out by matching tracker tracks with $p_T > 0.5$ GeV and total momentum $p > 2.5$ GeV to the muon system, where at least one muon segment present at a compatible position is required.

- **Global muon** tracks, built outside-in by matching standalone-muon tracks to tracker tracks after checking the compatibility of their parameters; the global fit is done on both tracks together. Due to the large size of the muon chambers, the combined fit improves the momentum resolution compared to the tracker-only fit for muons of $p_T > 200$ GeV.

The reconstructed muons undergo a set of identification criteria based on the quality of the tracks; these concern observables such as the track fit χ^2, the number of hits per track or the degree of matching between the tracker and standalone tracks. Different requirements result in different muon identification working points; three types of muon identification are defined based on efficiency and fake rate (rate of particles misidentified as muon): **Loose, Medium and Tight muon** identification correspond respectively to efficiencies of $\sim$ 99.7, 98.5, and 97% in data.

The information of the muon subdetectors is further analyzed to compute the isolation of the muon candidate, which assists in distinguishing between prompt muons and those originating from weak decays within jets. The PF relative isolation is defined as the ratio between the sum of all PF candidates' p_T within a cone size $R < 0.4$ around the muon and the p_T of the muon itself. Loosen the cutoffs to this value give six tight muon isolation working points.

In consequence of the muon system's upgrade and the algorithms' improvements (see Sections 2.2.1, 2.2.3), the muon reconstruction performance in Run 2 was comparable to Run 1, notwithstanding the increased instantaneous luminosity and pileup. The reconstructed hits' spatial resolution lay between 50 and 300 μm, leading to a muon selection efficiency above 96% [83]. Thanks to the strong CMS magnetic field and the excellent resolution of the muon system, the CMS detector is ideally suited to reconstruct dimuon candidates in a mass ranging up to 150 GeV, which was instrumental in the Higgs boson discovery.

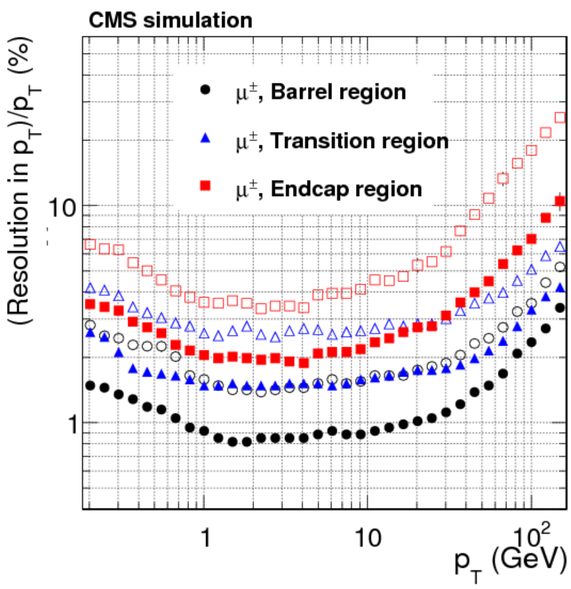


Figure 2.14: Muon transverse momentum resolution as a function of the p_T in the barrel, transition, and endcap regions for isolated muons. For each bin in p_T, the solid (open) marks correspond to the half-width for 68% (90%) intervals of the residuals distribution. [63]

The resolution of the transverse momentum of the muons is affected by two main sources of uncertainty: the first is the multiple scattering of the muons at the detector, and the second is the non-infinite precision of the hit positions. The effects of multiple scattering depend on the material budget of the detector; thus, the associated uncertainty depends on the pseudorapidity. The introduced deflection angle is proportional to $1/p_T$; therefore, the corresponding uncertainty is proportional to p_T. The error on the hit position introduces uncertainty proportional to p_T^2. The multiple scattering effects are the major source of uncertainty in measuring the transverse momentum for muons with $p_T < 30$ GeV, while the second effect is dominant for high-energy muons. The uncertainty of the transverse momentum is estimated in simulated events by calculating the residuals, which are defined as the differences between the generated and reconstructed tracks p_T. Fig. 2.14 shows the resolution of the transverse momentum uncertainty as a function of p_T of the muon. The half-widths of the residual distribution are computed for each bin, containing 68% and 90% of the events.

2.4.3 Electrons

The second particle to be reconstructed at CMS, after muons, is the electron. Electrons lose a significant fraction of their energy in the tracker, foremost due to the non-Gaussian phenomenon of bremsstrahlung. At $\eta \approx 0$, an electron loses an avarage of 33% of its energy and up to ∼86% at $\eta \approx 1.4$, where the highest material budget. Therefore their reconstruction is more complicated: the tracker algorithm must consider the energy loss, and the clustering algorithm must accumulate the bremsstrahlung photon energy, which can be positioned far away from the electron interaction point in the ECAL. The electron reconstruction overcomes these challenges with a dedicated tracking algorithm and an advanced energy clustering procedure.

The electron tracks in CMS are reconstructed as part of the PF reconstruction by matching tracker tracks with ECAL clusters to recover bremsstrahlung photons and potential photon conversions. The electron charge measurement is affected by bremsstrahlung, mainly when the bremsstrahlung photons convert upstream in the detector. The methods used widely to estimate the charge are the curvature of the electron track fitted with the Gaussian-Sum Filter (GSF) [84, 85] algorithm, the curvature of the KF track matched to the electron, and the different position of the ECAL cluster or supercluster with respect to the first hit of the GSF track. Then, the charge is chosen as the one given by at least two of these three measures. The charge misidentification probability of this algorithm is predicted by the simulation using electrons coming from Z boson decays without additional purity selections to be 1.5%.

The electron momentum is measured using both the ECAL calibrated energy and the momentum of the GSF track, which are combined using a multivariate regression aiming at the relative weight of the tracker and ECAL measurements. The resolution is optimal at the barrel region where low bremsstrahlung electrons (2%) while it ranges from 5-10% for showering endcap electrons of p_T up to 100 GeV, as shown in figure 2.15.

The identification of isolated electrons aims to separate them from jets misidentified as electrons and photon conversions. Multivariate classifiers performed via Boosted Decision Trees (BDT) are primarily used for this purpose. The variables used in the BDT training are observables that compare

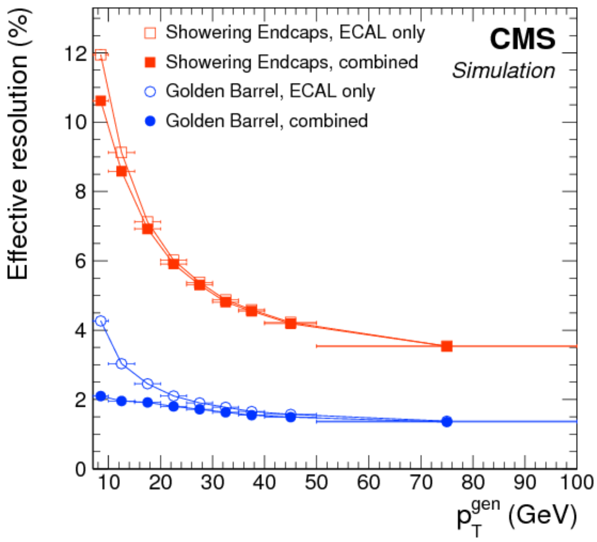


Figure 2.15: Resolution in electron momentum for the combined ECAL energy and track momentum, compared with the ECAL energy only, as a function of the generated electron p_T. Electrons in the barrel with the best resolution (Golden electrons) are shown in blue, and showering electrons in the endcaps with the worst resolution are shown in red. [85]

measurements obtained from the ECAL and the tracker, tracking only observables, and calorimetric only observables. Calorimetric observables profit from the transverse shape of the ECAL deposits, the energy in HCAL and the preshower to help reject jets with large electromagnetic components. Tracking observables enhance the separation between electrons and charged hadrons; they are based on the GSF and the KF tracks. Moreover, the isolation, which uses PF candidates for offline analysis, the hit pattern and impact parameters work as a discriminator against converted photons.

2.4.4 Photons

Most photons traverse the tracker without interacting and then deposit most of their energy ($\sim$ 97%) in the ECAL, where the shower can expand into various neighbouring crystals. Therefore, the photon reconstruction [86] is seeded from ECAL superclusters with $E_T > 10$ GeV, unrelated to any GSF track. These superclusters must be isolated from other ECAL clusters and the extrapolated tracks from the tracker, and the distribution of their energy deposit must be compatible with a photon shower. Sometimes, photons interact with the tracker material and convert into electron-positron pairs before penetrating the ECAL. Converted photons are distinguished because the electron-positron pair leave bent trajectories in the tracking system, with approximately identical momentum due to the photon being massless. The corresponding tracks can be seeded by displaced secondary vertices and extrapolated outward or by ECAL superclusters and extrapolated inwards. The standard energy resolution for unconverted (converted) photons with $p_T > 25$ GeV in the barrel is $\sim$ 1% (1.3 - 2.5%) and 2.5% (3%) in the endcap regions [86]. The high precision of the photon energy reconstruction enhances the selectivity of the $H \to \gamma\gamma$ decay, one of the key channels for the precise measurements of the Higgs boson properties.

Neutral mesons decaying to photons are considered the essential source of photon misidentification. Photons originating from mesons are typically collimated, usually reconstructed as a single one.

Reducing the contribution of this background depends profoundly on the isolation of the photon candidate, which is defined by using a cut-based approach on a set of discriminating variables or by applying multivariate techniques, in both cases using observables similar to the electron identification BDT.

2.4.5 Jets

Quarks and gluons originating from the hard interactions go through the hadronization process in the detector and produce several collimated particles grouped in cone-shaped jets. The momenta of the jets are estimated from the collected hadronic products, which can be charged or neutral. These hadrons are reconstructed with the remaining PF elements after reconstructing the electrons, muons and photons. Jets typically deposit energy in the ECAL and the HCAL; hence the clusters of both calorimeters are combined first. If the HCAL cluster is reconstructed within the tracker acceptance and is not linked to any other track and no ECAL cluster is found, it is identified as a neutral hadron. The remaining HCAL clusters are combined to one or several tracks, which may also be linked to ECAL clusters; they are reconstructed as charged hadrons.

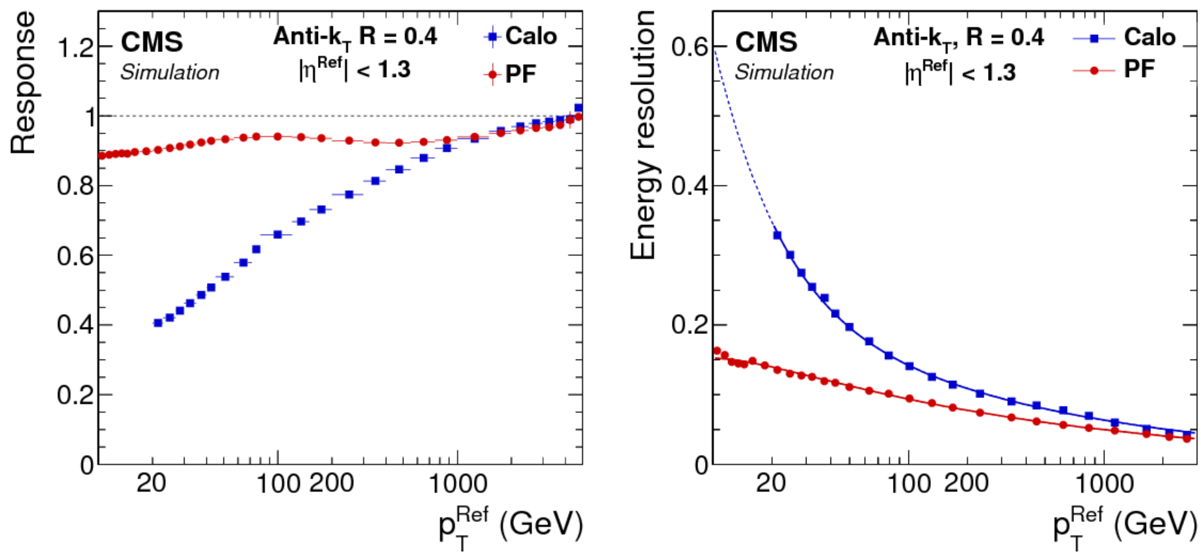


Figure 2.16: Jet energy response (left) represents the mean ratio between the reconstructed jet and the reference jet energy. Jet energy resolution (right) represents the ratio between the Gaussian width of the corrected and reference jet energies as a function of the reference jet p_T. Results are shown for jets in the barrel reconstructed with the anti-k_T algorithm with $\Delta R = 0.4$ following the PF and calorimeter-based approaches. [81]

The reconstruction of Jets is done by clustering charged and neutral hadrons with the anti-k_T algorithm [87]. The clustering is performed recursively by matching the PF candidates, which are close to each other, according to a metric that produces jets of conic shape. The size of the cone is calculated by the distance parameter that can be set to 0.4 or 0.8. In both cases, the pairs with the highest p_T are clustered first; this way, the cone is built around the hardest particle of the event, which suppresses soft radiation and collinear parton splitting at the borders. The jet four-momentum is estimated as the vector sum of the clustered PF candidates four momenta; it differs from the original parton due to theoretical uncertainties on the hadronization process and other experimental effects.

A set of jet energy corrections (JEC) is applied as a function of (p_T, η) to calibrate the jet response taking into account the following: pileup, detector noise, the response, non-linearity and inhomogeneity of the calorimeters. The typical reconstructed jet energy has a response of $\sim$ 0.9 and a resolution lower than 15% for all p_T ranges, as shown in Fig. 2.16, where the improvement of the PF approach is noticeable with respect to calorimeter-based reconstruction. The jets are further subject to identification requirements due to instrumental noise in the calorimeters and suppress those poorly reconstructed. Two different levels of selection are defined (loose, tight) based on the fraction of charged and neutral hadrons in the jet, the fraction of energy deposited in the ECAL, and the charged hadron multiplicity.

2.4.5.1 Quark-gluon Jet Discrimination

Quark and the gluon-induced jets can be distinguished in CMS with the Quark-Gluon Likelihood [88] algorithm; it exploits the fact that gluon-induced jets are wider with higher multiplicities and contain a more uniform energy fragmentation. In contrast, quark jets are more likely to produce narrow jets with hard constituents that carry a significant fraction of the energy. These differences can be used to build a probability tagger that discriminates jets initiated by quarks from those initiated by gluons.

The quark-gluon likelihood discriminant is constructed out of three variables, each accessing the particle flow composition of the jet:

- **The multiplicity**:
  The jet constituents multiplicity represents the total number of particle flow candidates reconstructed within the jet. The quark-jets tend to have a smaller number of constituents with respect to gluon-jets. This variable is particularly discriminating for high p_T jets where the average quark-to-gluon multiplicity ratio would converge to 4/9 [89].

- **The jet energy sharing variable:**
  Quarks fragmentation produces hard constituents that carry a significant fraction of the jet energy. The jet fragmentation distribution is defined by

  $$p_T D = \frac{\sqrt{\sum_i p_{T,i}^2}}{\sum_i p_{T,i}} \tag{2.10}$$

  which has $p_T D \to 1$ for jets made of only one particle that carries all of its momentum and $p_T D \to 0$ for a jet made of an infinite number of particles.

- **The angular spread**
  The angular spread of jets is measured by a minor axis σ_2 of the jet in the η - ϕ plane. The p_T-weighted shape of the jet in the η - ϕ plane is approximated by an ellipse. The ellipse is identified by the minor and a major axis length and orientation. The minor axis length is used to calculate the quark-gluon likelihood, and it is particularly beneficial for low p_T jets. In order to find the minor axis length, the p_T-weighted axis of the jet is estimated, and the

2×2 matrix M is constructed as:

$$M_{11} = \sum_i p_{T,i}^2 \Delta\eta_i^2 \quad (2.11)$$

$$M_{22} = \sum_i p_{T,i}^2 \Delta\phi_i^2 \quad (2.12)$$

$$M_{12} = M_{12} = \sum_i p_{T,i}^2 \Delta\phi_i \Delta\eta_i \quad (2.13)$$

where the sum is performed all over the constituents, $p_{T,i}$ is the transverse momentum. $\Delta\eta_i$ and $\Delta\phi_i$ are the pseudorapidity and azimuthal distance between the constituent and the jet axis.

Let λ_1 and λ_2 be the major and the minor eigenvalues of the matrix M, the p_T-weighted minor axis length is

$$\sigma_2 = \sqrt{\frac{\lambda_2}{\sum_i p_{T,i}^2}} \quad (2.14)$$

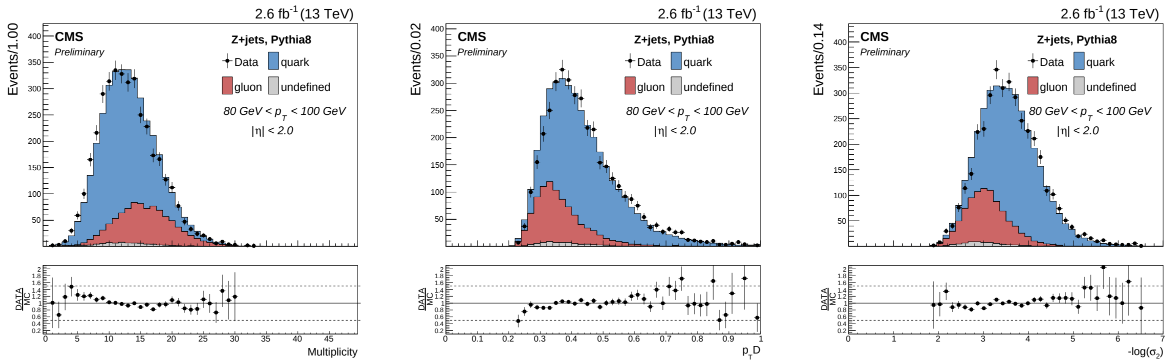


Figure 2.17: The quark-gluon likelihood input variables distribution in Z+jets events for jets with 80 < p_T < 100 GeV and $|\eta| < 2$. The jet constituents multiplicity (left), The jet energy sharing (middle), and the angular spread (right). [90]

Fig. 2.17 shows the distributions of the three described variables in Z+jets events for jets with 80 < p_T < 100 GeV and $|\eta| < 2$.

The set of variables used in the discriminator have been chosen by studying their discrimination performance on the simulation. Their discrimination power is compared with receiver operating characteristic (ROC) curves. These curves show the efficiency of selecting a quark jet for a selection cut on the variable that rejects a given fraction of gluon jets. In addition to the ROC curves, correlations are studied to obtain the minimal set of uncorrelated variables with the highest discrimination power. A better discrimination power and stability to pileup effects are found by restricting the charged particle flow candidates to those linked to tracks compatible with the primary interaction vertex and restricting the neutral particle flow candidates to those which have $p_T > 1$ GeV.

The quark-gluon likelihood has been validated using the Z+jet sample as enriched quark events and dijet sample, which are gluon enriched. In addition, data-driven scale factors have been computed

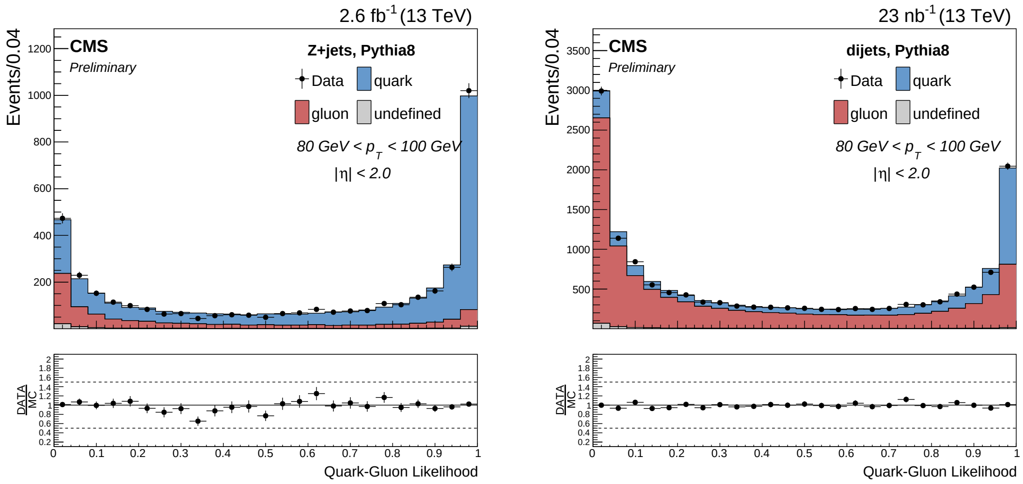


Figure 2.18: The quark-gluon likelihood discriminant for jets with $80 < p_T < 100$ GeV and $|\eta| < 2$ in Z+jets (left) and dijet (right) events. [90]

and fitted with polynomial functions to be applied separately for quarks and gluon jets in MC events. Fig. 2.18 shows the quark-gluon likelihood output distribution in data and MC for the two enriched samples. The MC events include the polynomial corrections, while uncorrected events have been used to estimate the systematic uncertainty.

2.5 Simulation

The simulation is an essential aspect of high-energy physics experiments. It provides approaches to predict distributions based on theoretical assumptions; these distributions can be directly compared with the output of real-world experiments, such as CMS. This section gives a short description of the simulation techniques used in the analysis documented in this thesis.

2.5.1 Event generators

Proton-proton collisions are very complex and challenging to model accurately. Different techniques are used to describe the QCD processes, whose phenomenology varies significantly at different energy scales. Protons are composed of 3 quarks (valence quarks), virtual gluons, and virtual quark anti-quark pairs coming from gluon splitting. All constituents of hadrons are called partons.

During high energy collisions, the hard scattering is described at the level of parton interactions. The hadronic cross-section σ_{pp} is calculated based on the QCD factorization theorem. The factorization theorem states that the hadronic cross-section σ_{pp} is a convolution of the partonic cross-section $\hat{\sigma}_{i,j}$ with the parton distribution functions (PDFs) $f_i(x)$:

$$\sigma_{pp} = \int_{x_{min}}^{1} \sum_{i,j} f_i(x_1) f_j(x_2) \hat{\sigma}_{i,j}(x_1 p_1, x_2 p_2) dx_1 dx_2 \tag{2.15}$$

where function $f_i(x)$ is probability density that a parton of type i has a fraction x of the hadron energy.

In addition to the hard interaction, the other proton constituents can also interact, resulting in a spray of softer particles, called the underlying event. Additional hard QCD radiation will be emitted from the high momentum particles involved in the collision. Before the hard interaction, the radiation that results from particles is called initial-state-radiation (ISR), while radiation off particles produced in the collision is called final-state-radiation (FSR). Quarks and gluons can emit additional radiation through strong interaction. All the quarks and gluons pass through the hadronization process, forming hadrons. Eventually, unstable particles are going to decay. A representation of all these processes is presented in Fig. 2.19.

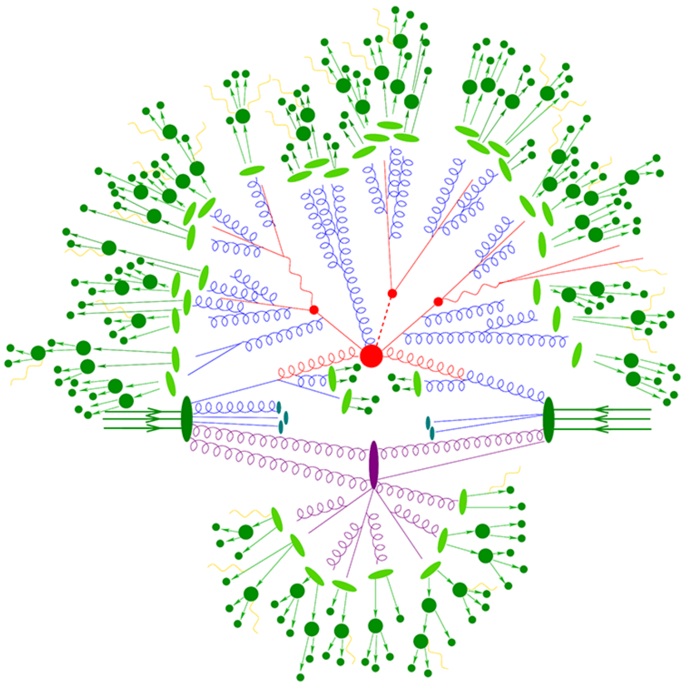


Figure 2.19: Representation of a proton-proton collision event as produced by an event generator. The hard interaction (big red blob) is followed by the decay of the products (small red blobs). Initial and final state radiation is represented in blue. A secondary interaction (in purple) can take place before the final-state partons hadronization. The hadronization is shown by the green blobs and hadron decay in dark green. Photon radiation is presented in yellow. [91]

A simulation sequence of such a process could be done using several software packages in the following steps:

- **Matrix Element (ME)**: responsible for calculating the scattering amplitude utilizing the proton PDFs to inspect the initial state particles' momenta and produce the final state particles [92].
- **Parton Shower (PS)**: used to simulate additional emissions in perturbative QCD. The probability of producing an event with a given final state is approximately similar to the probability of the same final states in the real-world experiment [93].
- **The hadronization**: describe the transition from colored particles to hadrons, is treated using phenomenological models [94].
- **Underlying Events**: the decay of unstable particles, modeled based on experimental data [95].

The mentioned steps are done using software packages called event generators. The event generator programs used in this analysis are PYTHIA [96], MADGRAPH [97], and POWHEG [98].

2.5.2 Detector simulation

After completing the MC event generators' whole chain, Monte Carlo techniques are then used to simulate the interaction of the particles with the detector. The detector simulation is implemented using GEANT4 [99]. Pileup interactions are also added at this stage: according to a specified pileup scenario, a library of simulated hits of minimum bias events adds extra interactions to the signal event. Once the detector simulation is operated, all the detector signals are converted to electronic signals in a format identical to the one used for data. After this step, the simulated events go through the same reconstruction steps as the collision data.

Chapter 3

CMS track reconstruction performance during Run 2

The precise and efficient measurement of charged particle momentum is a critical component of CMS's physics program, as it impacts the ability to reconstruct the physics objects needed to understand pp collisions at the LHC. With the start of Run 2, the LHC has resumed delivering pp collisions with a center-of-mass energy of 13 TeV and a bunch time separation of 25 ns, with an average of more than 25 inelastic collisions superimposed on the event of interest. In this highly populated environment, accurate reconstruction of the particle momentum is one of the major challenges of CMS. In this chapter, the CMS tracking strategy used for Run 2 is described. In addition, the performance of the muon tracking measured with the tag and probe technique on the $Z \longrightarrow \mu^+\mu^-$ di-muon resonances in the legacy data of Run 2 at $\sqrt{s} = 13$ TeV is discussed. The results presented in this chapter were published in [100].

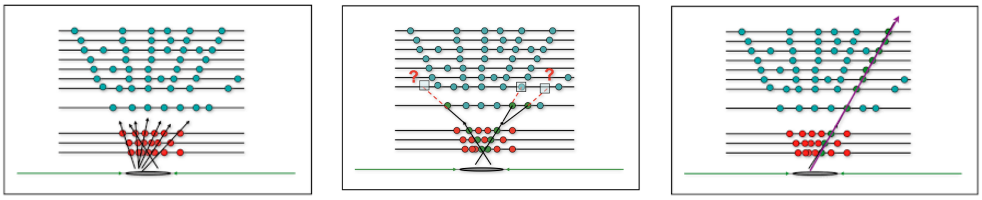
3.1 Track reconstruction

A tracker hit is the best estimate of the point where a charged particle has crossed the silicon layer, based on the released charge distribution. Each trajectory is a sequence of hits in the tracker: tracking algorithms assign hits to tracks and aim to measure the five parameters of the helix with the best possible resolution.

The CMS track reconstruction algorithm [63] is following an iterative approach: the track reconstruction is run several times, the first set of iterations looks for tracks that are easier to find, i.e. non-displaced tracks and with relatively high p_T, and progressively proceeding to the more complex ones. After each iteration, hits associated with reconstructed tracks are then removed from the set of hits before running the following iterations. This procedure reduces the combinatorial complexity so that the more difficult kinematic regions tracks can be reconstructed.

The algorithm used for the track reconstruction is called the Combinatorial Track Finder (CTF), which is based on the combinatorial Kalman filter (CKF) technique [82]. The CTF algorithm can be described in a 4-step procedure that is run at each iteration:

- **Seed generation (seeding)**: It provides the initial estimate of the trajectory parameters and their uncertainties (see Fig. 3.1 (a)). The seeding needs at least three hits (or two with



(a) Seeding. **(b)** Building. **(c)** Fitting.

Figure 3.1: Track reconstruction in CMS [101]

an additional constraint on the track's origin). Pixel hits are used for seeding because they are 3D hits with better resolution and no matching ambiguity as for the strips. The new pixel detector with four layers, described in Section 2.2.1, improved the seeding significantly: the best seeds now have 4-pixel hits.

- **Pattern recognition (building)**: The proto-candidates are then propagated, searching for compatible hits at each layer using the Kalman filter. The seeds are extrapolated to the entire tracker, and at each iteration, the track parameters are updated. Once the track is complete, another search is performed backward, starting with the outermost hit. This step is performed because the pattern recognition is more efficient than the seeding, as it correctly handles possible overlaps of silicon modules and can recover pixel hits that were not used at the time of seeding. If multiple compatible hits are found when extrapolating the trajectory to a single layer, the algorithm creates a candidate for each hit, propagating independently. Ultimately, only one trajectory is kept based on the total number of hits and the quality.

- **Track fitting**: Once the tracks are built, the trajectory is fitted after combining all the associated hits using an inside-out and outside-in Kalman filter, with the last step, also called the "smoothing stage". The Kalman filter is firstly run starting from the innermost hit, where the state of the trajectory is given by the inner hits. For each valid hit, the estimated uncertainty of the hit position is also updated using the values of the track parameters. The smoothing stage follows this first filter: a second filter is initialized with the result of the first filter but with enlarged errors and is run outside-in. The helix parameters can be obtained from the weighted average of the parameters of these two filters at the surface associated with each hit.

- **Track selection**: Tracks are selected based on quality flags: the number of hits in the entire tracker, the number of 3D hits, the distance from the primary vertex, and the track normalized χ^2.

Iteration	Seeding	Target tracks
Initial	pixel quadruplets	prompt, high p_T
LowPtQuad	pixel quadruplets	prompt, low p_T
HighPtTriplet	pixel triplets	prompt, high p_T recovery
LowPtTriplet	pixel triplets	prompt, low p_T recovery
DetachedQuad	pixel quadruplets	displaced−−
DetachedTriplet	pixel triplets	displaced−− recovery
MixedTriplet	pixel+srtip triplets	displaced−
PixelLess	inner srtip triplets	displaced+
TobTec	outer srtip triplets	displaced++
JetCore	pixel pairs in jets	high-p_T jets
Muon Inside-Out	muon-tagged tracks	muons
Muon Outside-In	standalone muons	muons

Table 3.1: Summary of the seeding hits and the target tracks for each of the tracking iterations.

The contribution of each iteration to the final reconstructed tracks is shown in Fig. 3.2 as a function of p_T, with several tracking iterations in different colors. Simulated events with the new pixel detector $t\bar{t}$ production with an average pileup of 35 have been used. The plot shows that the tracking efficiency saturates at $\sim$ 90% due to hadrons that undergo nuclear interactions in the tracker material, while it is known to be close to 100% for muons. At low p_T, the efficiency is lower due to the nuclear interactions. The different tracking iterations are also listed in Table 3.1 with their seeding hits and target track: the first set of iterations uses triplets with a compatible fourth hit, and the second propagates the triplet. In the subsequent iterations, a 3-hit seeding is run as a recovery for prompt tracks, then displaced tracks with several combinations of hits are considered. The final steps develop special iterations to improve tracks reconstruction in a high-density environment (jets) or use information from other subsystems (muons).

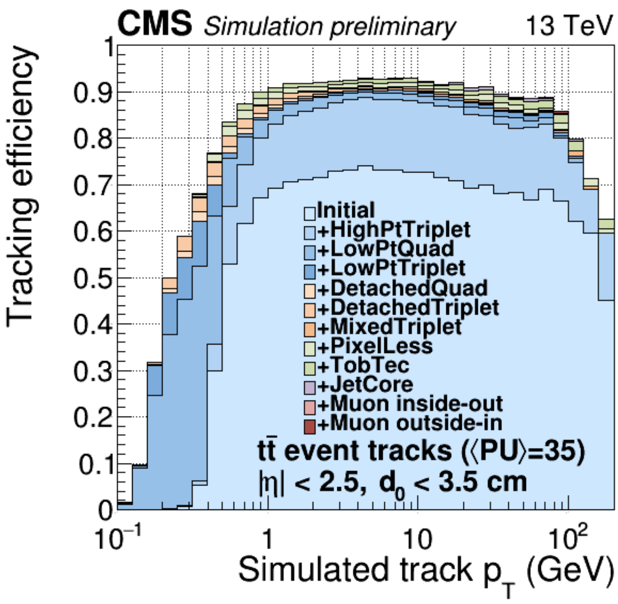


Figure 3.2: Iterative tracking efficiency in MC as a function of track p_T. [102]

At the end of the iterative process, the tracks that fulfill some minimum requirements are kept in the *"All-Tracks"* collection. The first set of iterations are only seeded by hits in the inner tracker, while the last two steps use muon candidates from the muon system to create seeds for the track reconstruction in the inner tracker, reconstructing muon-tagged tracks. Tracks that do not belong to these two specific iterations are kept in the so-called *"Tracker-only seeded tracks"* collection and are particularly important for studies at low transverse momentum or pion reconstruction. Both *"Tracker-only seeded tracks"* and *"All-Tracks"* collections have been used extensively for muon-track efficiency studies mentioned later in this context.

3.2 Datasets

3.2.1 Data

In late 2015 and early 2016, the Silicon Strip Tracker observed a decrease in signal-to-noise, also associated with a loss of hits on tracks. This behavior was due to saturation effects in the pre-amplifier of the APV25 readout chip [103]. The drain speed of the pre-amplifier was affected more strongly by the change in operating temperature than anticipated, leading to a prolonged discharge of the amplifier under high occupancy conditions. About 20 fb^{-1} of 2016 data were affected by this issue, referred to as (old APV settings). The drain speed was changed to allow for faster recovery. With the new APV pre-amplifier settings, the hit efficiency is back at the same level as in Run 1, 16 fb^{-1} of 2016 data has been taken with the new settings, referred to as (new APV settings).

This analysis uses proton-proton collision Legacy data recorded by the CMS experiment during Run 2 at $\sqrt{s} = 13$ TeV, corresponding to an integrated luminosity of 59 fb^{-1} (2018), 41 fb^{-1} (2017), 20 fb^{-1} (old APV settings), and 16 fb^{-1} (new APV settings) and based on a single muon trigger.

3.2.2 Simulation

In order to improve the description of 2016 data in the period affected by the APV saturation issue, around half of the Monte Carlo events generated for the 2016 Run 2 Legacy campaign underwent a dedicated simulation of the APV25 chip dynamic gain in the Strip Tracker partitions. A set of APV baseline distributions, representing the charge accumulated on a strip over several bunch crossings, are provided as an input, separately per layer, and in bins of PU and z within a layer. Each time a charge is deposited on a strip, the APV response is simulated; this is linear for low APV baselines (small amount of charge from previous interactions remaining on strips) but becomes non-linear for high APV baselines; the net effect is that the charge effectively deposited/readout from a strip is reduced.

Drell-Yan + jets samples generated using MadGraph5 generator at LO QCD accuracy are used in this analysis. Events are weighted to match the pile-up distribution in the data.

3.3 The Tag and Probe method

The calculation of detector efficiency is an essential aspect for any physics analysis. It considers the particles generated in the collision and escaped detection (dropped by the reconstructions algorithms, did not arrive the detector components, etc.). It could be evaluated from simulation, but simulation needs to be calibrated with data. The tag and probe method (T&P) is a data-driven technique used in CMS since 2008 to measure the efficiencies from data, with no reference to simulation. It is based on the reconstruction of well-known resonances, such as J/ψ or Z.

In this method, resonances used to calculate the detector efficiencies decay to a pair of particles, which in our case are muons, One of which has passed a tight identification ("tag") while the other one has passed a loose identification ("probe"). Then, the "probes" are divided into two categories, i.e. "passing" and "failing" probes, depending on specific selection criteria. Di-muon candidates are produced by pairing tag muons with probe muons, requiring the invariant mass of the pair to be close to the resonance mass and the total charge to be zero.

Some procedure is applied at the tag+probe invariant mass spectra for the passing and failing probes to subtract the non-resonant background (e.g. from fake probes) to calculate the net number of passing and failing di-muon probes, as described in section 3.3.2. Finally, the efficiency is computed as the ratio between the "passing probes" and the total number of probes in the sample. A more detailed description of this method is given in Ref. [104].

This chapter is focused on the extraction of the muon tracking efficiency using the tag and probe technique using $Z \longrightarrow \mu^+\mu^-$ di-muon resonances. Assuming that the muon reconstructions in the tracker and the muon system are independent, the efficiency is given by the probability that a track reconstructed in the inner silicon tracker is associated with a given standalone muon "probe". The generic definition of "probe" and "tag" muons are similar to the ones already used for RunI data [105]:

- **tag:** A global muon with transverse momentum p_T larger than 27/29 GeV [1], geometrically matched to a trigger object that fired the single muon trigger for isolated muons with a nominal p_T threshold of 24/27 GeV [2]. In addition, relative combined isolation with $\Delta\beta$ correction [73] in $\Delta R = 0.4$ is applied to be less than 0.15.
- **probe:** Any standalone muon with at least one valid hit in the muon system (i.e. good track-hit χ^2).
- **passing probe:** The standalone muon is matched with tracks that fulfills minimum quality requirements in ($\Delta R < 0.3$). The matching is defined by comparing the directions at the point of the closest approach to the beamline of the two tracks. The matching will be discussed in details in section 3.3.1.

3.3.1 Matching

There are two challenges in measuring the tracking efficiency based on the matching between the two subdetectors: the possibility that the standalone muon is linked with a spurious tracker track when the true one is not reconstructed and the probability that the track which is successfully reconstructed in the inner tracker is not associated with a standalone muon. The former possibility

[1] $p_T > 27$ GeV(2018, 2016 dataset) and $p_T > 29$ GeV (2017 dataset).
[2] single muon trigger with a nominal p_T threshold of 24 GeV (2018, 2016) and 27 GeV (2017 dataset).

leads to overestimating the efficiency and becomes more significant if the matching criteria are made looser, while the latter possibility can lead to underestimating the efficiency and can be mitigated by choosing very loose matching criteria. The two effects must be controlled to measure the tracking efficiency with high accuracy or if large deviations between simulations and real data are expected.

The measured efficiency ϵ can be formulated in terms of the true tracking efficiency ϵ_T, matching efficiency ϵ_M and the the probability of spurious matches ϵ_F as [105]

$$\epsilon = \epsilon_T \epsilon_M + (1 - \epsilon_T \epsilon_M)\epsilon_F \tag{3.1}$$

As it will be described in section. 3.4.1, ϵ_F can be measured from data. Once we have ϵ_F, one can invert the relation and obtain

$$\epsilon_T \epsilon_M = \frac{\epsilon - \epsilon_F}{1 - \epsilon_F} = 1 - \frac{1 - \epsilon}{1 - \epsilon_F} \tag{3.2}$$

The true inefficiency equals the measured inefficiency scaled up by a factor $1/(1 - \epsilon_F)$. The uncertainty on the corrected efficiency can be computed as

$$\sigma(\epsilon_M \epsilon_T) = \frac{1}{1 - \epsilon_F}\sigma(\epsilon) \oplus \frac{1 - \epsilon}{(1 - \epsilon_F)^2}\sigma(\epsilon_F) \tag{3.3}$$

Hence the non-negligible ϵ_F decrease the sensitivity of the measurement, as expected, while the uncertainty on ϵ_F has a small contribution to the precision of the measurement due to the suppression factor $(1 - \epsilon) \ll 1$
Extracting ϵ_T from $\epsilon_M \epsilon_T$ can be more problematic. The value of ϵ_M can only be extracted from simulation, not from data, and it is sensitive to the tails of the standalone muon resolution. There is no suppression factor reducing the contribution of $\sigma(\epsilon_M)$ so any small uncertainty on ϵ_M propagates directly to ϵ_T

$$\sigma(\epsilon_M) = \frac{1}{\epsilon_M}\sigma(\epsilon_M \epsilon_T) \oplus \frac{\epsilon_T}{\epsilon_M^2}\sigma(\epsilon_M) \tag{3.4}$$

The matching efficiency cannot be above 100%, so any measurement of the efficiency giving a confidence interval $\epsilon_M \epsilon_T \in [\epsilon_{low}, \epsilon_{high}]$ can be translated conservatively in a larger confidence interval for ϵ_T, $\epsilon_T \in [\epsilon_{low}, 100\%]$, this is usually enough, as the ϵ_T is expected to be very close to 100%. Also, the matching efficiency cancels out at least partially when computing the ratio of the tracking efficiencies between data and simulations.

3.3.2 Signal extraction

For the purpose of separating the signal (muons coming from the decay of true Z candidates) and the non-resonant background (e.g., from fake probes), a simultaneous fit is performed on the invariant mass spectra of the pairs for both passing and failing probes. This procedure is applied after splitting the data into sub-samples, corresponding to bins of the probe muon's kinematic variable of interest such as p_T, η, ϕ, and the efficiency will be measured as a function of that variable. Each sub-sample contains signal and background events; the signal is accessed by fitting the invariant mass spectra. The fit for each bin allows discriminating between signal and background statistically. In particular, the fit yields the number of signal events. The Z invariant mass distribution (signal) is modeled as Voigtian, while the background is modeled as an exponential function. The efficiency is finally calculated simply from the ratio between the signal yield from the fit of passing probes and all probes. This approach is illustrated in Fig. 3.3.

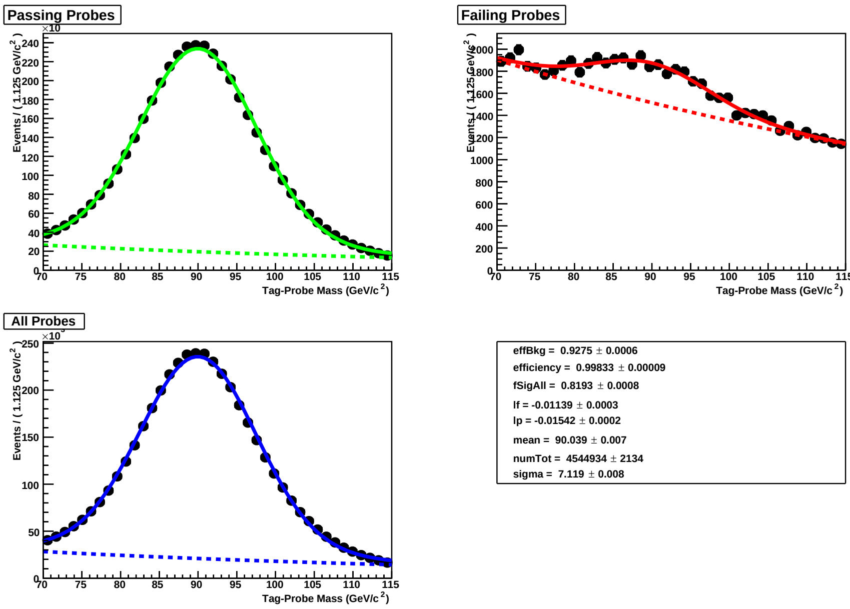


Figure 3.3: Invariant mass distributions for tag + passing probes (top-left), tag + failing probes (top-right), and tag + all probes (bottom-left). The Z mass window is selected to be in the [70-115] GeV range. Data is represented by black dots, and the fitted curves are also shown for each distribution (colored curves).

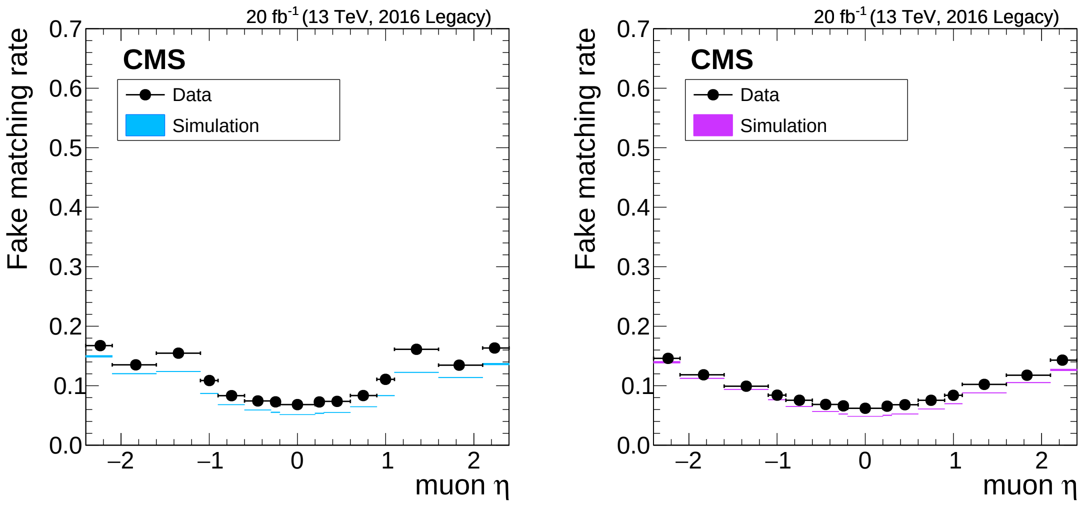


Figure 3.4: Fake matching rate as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "Tracker-only seeded tracks" (right), for 2016 Legacy (old APV settings). The uncertainties shown are statistical.

3.4 Results of the efficiency measurements

The muon tracking efficiency has been measured for "All-Tracks" and "Tracker-only seeded tracks" collections as a function of pseudorapidity η, azimuthal angle ϕ of the probe muon and the number of primary vertexes reconstructed in the event. The full results are available in [106], while the efficiency as a function of η only will be presented in this context.

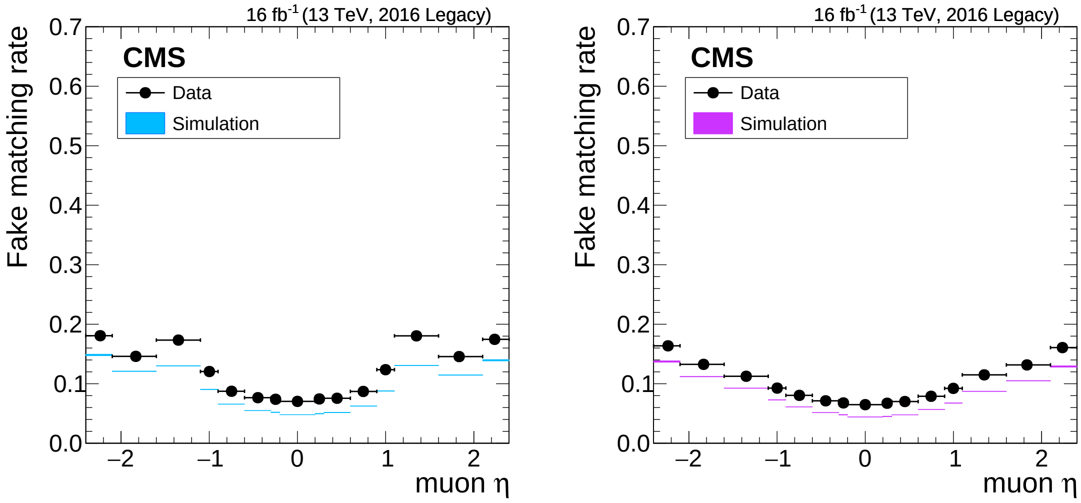


Figure 3.5: Fake matching rate as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "Tracker-only seeded tracks" (right), for 2016 Legacy (new APV settings). The uncertainties shown are statistical.

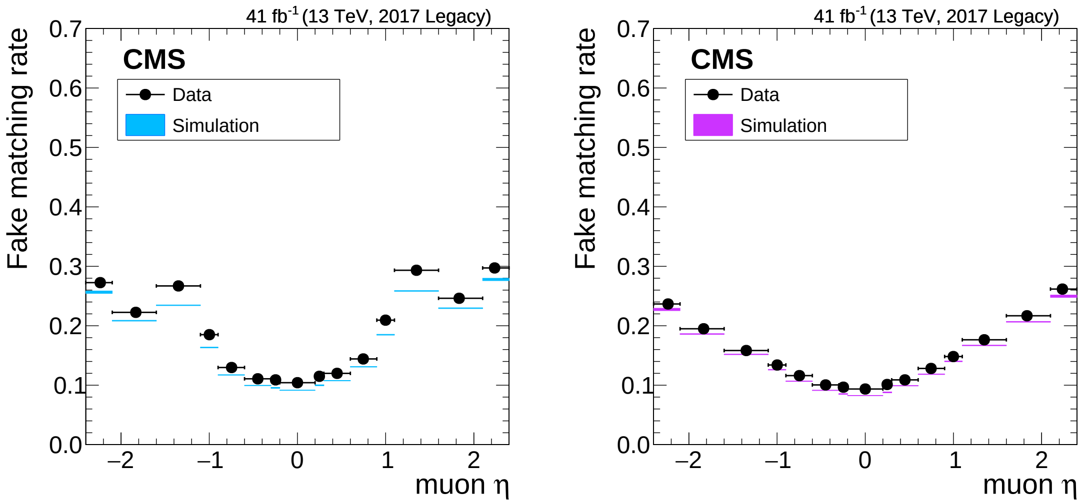


Figure 3.6: Fake matching rate as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "Tracker-only seeded tracks" (right), for 2017 Legacy. The uncertainties shown are statistical.

3.4.1 Measurements of the probability of spurious matches

As the matching criteria are very loose, a probe can also be matched to another track in the event, leading to an overestimate of the measured tracking efficiency, as explained in section 3.3.1. The probability of spurious matches can be measured directly on data: before matching probes and tracker tracks, one can remove those tracks that combined with the tag give an invariant mass near the Z peak. Under these circumstances, the tracking efficiency measured estimates the probability of fake matches.

The measured fake matching rate as a function of pseudorapidity η for "All-Tracks" and "Tracker-only seeded tracks" collections are presented in Fig. 3.4, 3.5, 3.6, 3.7 for 2016 (old APV settings), 2016 (new APV settings), 2017, and 2018, respectively. As expected, the fake match probability is small and becomes more visible in the endcap region due to the increased rate of hadrons above the momentum threshold used. The simulation reproduces the behaviour with a good agreement with data.

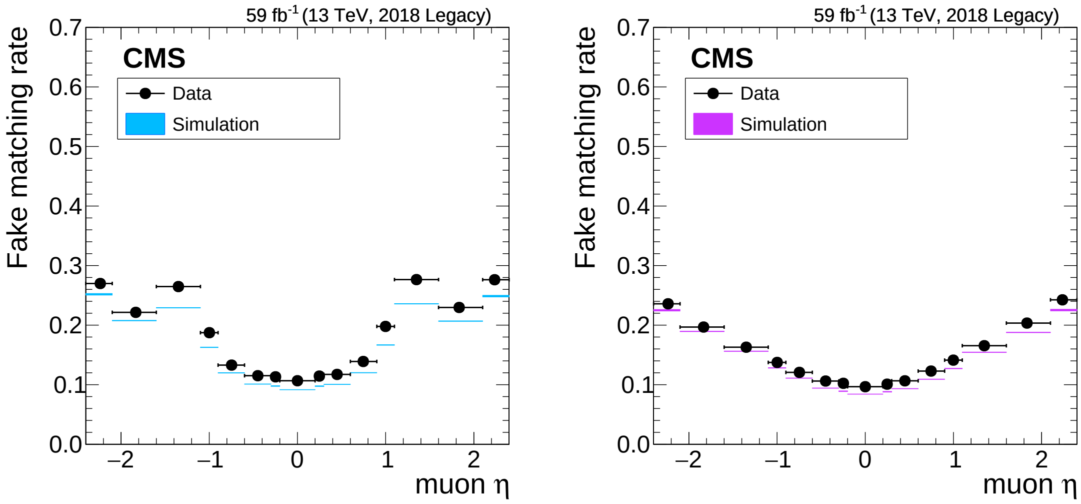


Figure 3.7: Fake matching rate as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "Tracker-only seeded tracks" (right), for 2018 Legacy. The uncertainties shown are statistical.

3.4.2 Corrected efficiencies

The measured efficiencies have been corrected according to equations 3.2, 3.3 using the estimate of the probability of fake matches shown in section 3.4.1.

The muon tracking efficiency for the 2016 (old APV setting) dataset with integrated luminosity 20 fb^{-1} has been shown in Fig. 3.8 as a function of pseudorapidity η of standalone muons. The efficiency using tracks in the "All-Tracks" collection (on the left) shows a good agreement between data and MC with a slight decrease in efficiency affected by the issue in the pre-amplifier of the APV25 readout chip, explained in section. 3.2. The effect of this issue is more pronounced in the case of the "tracker-only seeded tracks" (on the right), where there is a significant decrease in the efficiency with marked discrepancies between data and MC.

With the new APV pre-amplifier settings, the hit efficiency is back at the same level as in Run 1 [63]. The efficiency for 2016 taken with the new APV settings with integrated luminosity 16 fb^{-1} is shown in Fig. 3.9 as a function of pseudorapidity η of standalone muons. The efficiency using tracks in the "All-Tracks" collection (on the left) returned to its nominal value ($> 99.5\%$). The inefficiency in the "tracker-only seeded tracks", due to APV pre-amplifier saturated, around 4% has been eliminated with the new APV settings; the efficiency returned to its nominal value with a good agreement between data and MC. The biggest difference between data and MC $\sim 1\%$ is seen at the transition region between the barrel and endcaps.

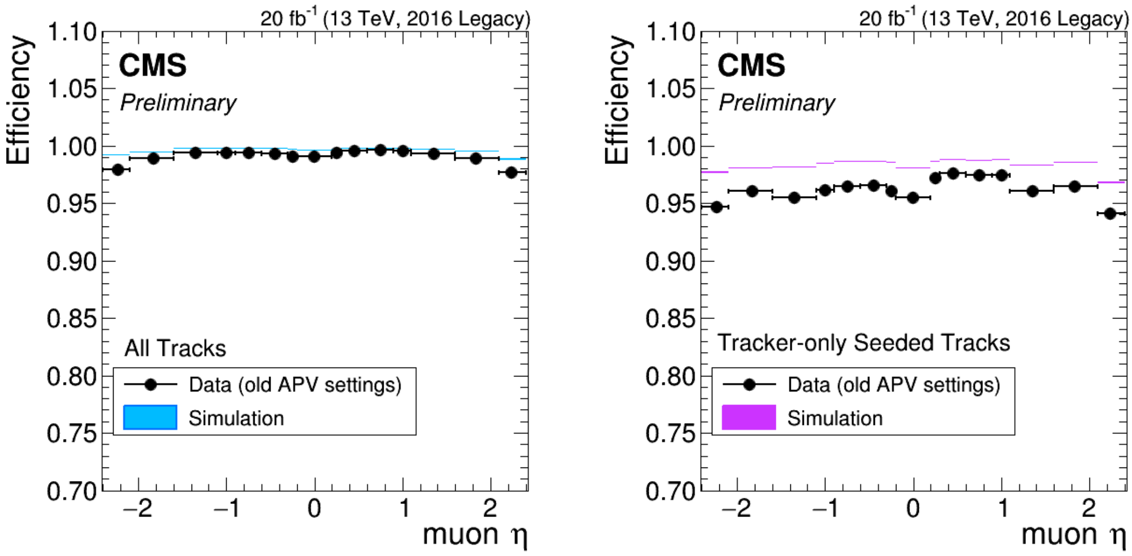


Figure 3.8: The muon tracking efficiency as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "tracker-only seeded tracks" (right). The 2016 (old APV setting) data is represented by black dots and simulation by colored rectangles. The uncertainties shown are statistical.

The muon tracking efficiency for both 2017 and 2018 datasets are shown in Fig. 3.10 and Fig. 3.11 respectively as a function of pseudorapidity η of standalone muons. The results obtained for the two years are so similar. The efficiency using tracks in the "All-Tracks" collection (on the left) shows a good agreement between data and MC with efficiency higher than 99.5% all over η regions. For "tracker-only seeded tracks" (on the right), the efficiency is higher than 97%, decreased to $\sim$ 95% in the transition region between the barrel and endcaps with the biggest difference between data and MC $\sim$ 1.5%.
Fig. 3.12 shows the tracking efficiency for Run 2 Legacy data and simulation as a function of pseudorapidity η of standalone muons. The efficiency using tracks in the "All-Tracks" collection (on the left) shows a good agreement between data and MC with efficiency higher than 99% all over η regions. For "tracker-only seeded tracks" (on the right), the efficiency is higher than 97%, decreased to $\sim$ 96% in the transition region between the barrel and endcaps with the most significant difference between data and MC $\sim$ 2%.

A comparison of the tracking efficiency for the different running periods of Run-2 is presented in Fig. 3.13. For "All-Tracks" (on the left), a good agreement of the tracking efficiency is observed for the running periods (2018, 2017, 2016 new APV settings), while for 2016 old APV settings, a loss in efficiency of up to $\sim$ 2% is found. For "tracker-only seeded tracks" (on the right), there is almost agreement of the efficiency observed for the running periods (2018, 2017, 2016 new APV settings), and a loss in efficiency of up to $\sim$ 5% is found for 2016 old APV settings.

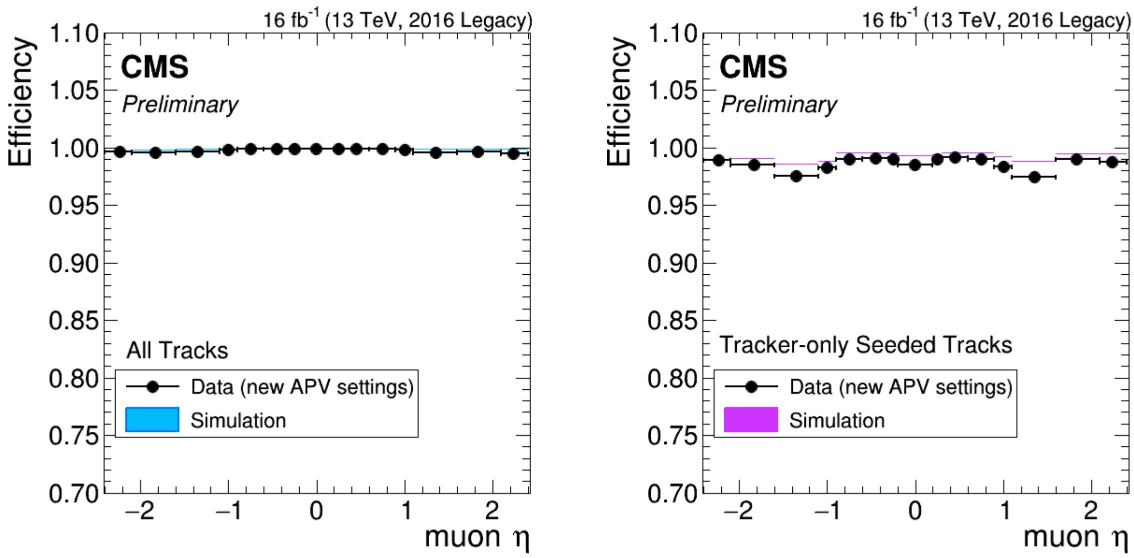


Figure 3.9: The muon tracking efficiency as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "tracker-only seeded tracks" (right). The 2016 (new APV setting) data is represented by black dots and simulation by colored rectangles. The uncertainties shown are statistical.

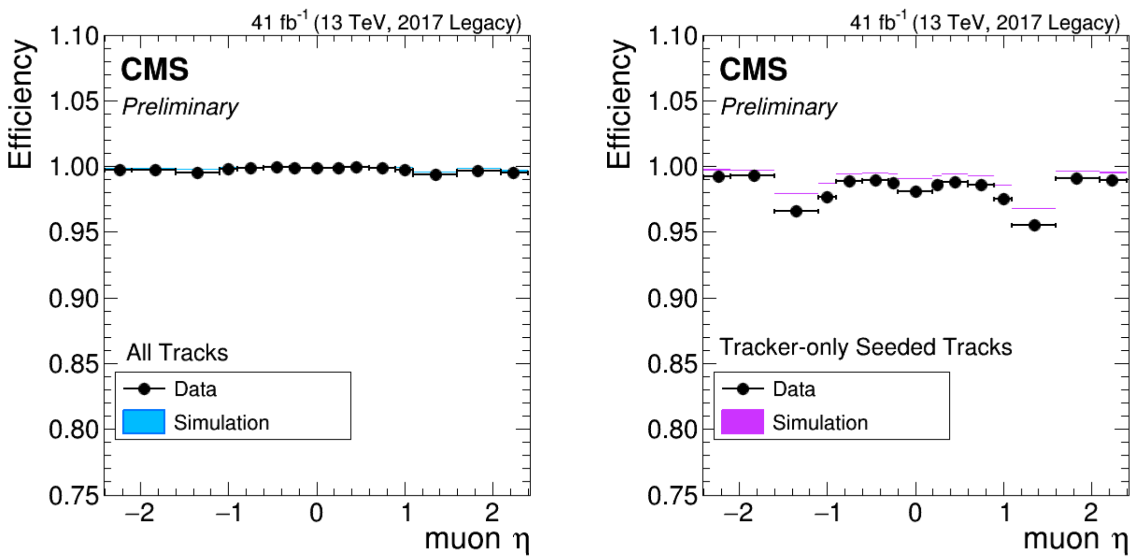


Figure 3.10: The muon tracking efficiency as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "tracker-only seeded tracks" (right). The 2017 Legacy data is represented by black dots and simulation by colored rectangles. The uncertainties shown are statistical.

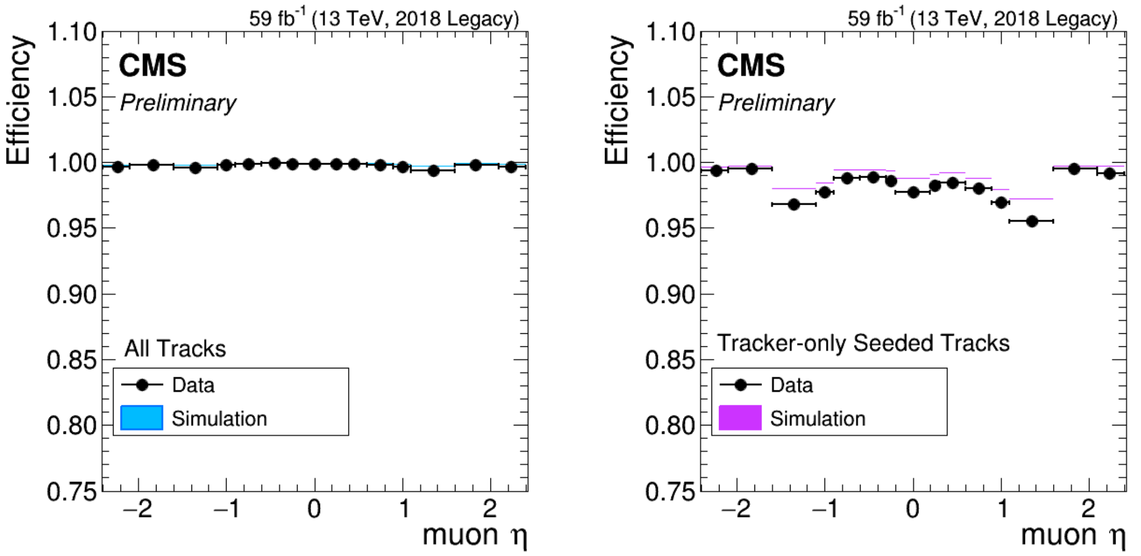


Figure 3.11: The muon tracking efficiency as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "tracker-only seeded tracks" (right). The 2018 Legacy data is represented by black dots and simulation by colored rectangles. The uncertainties shown are statistical.

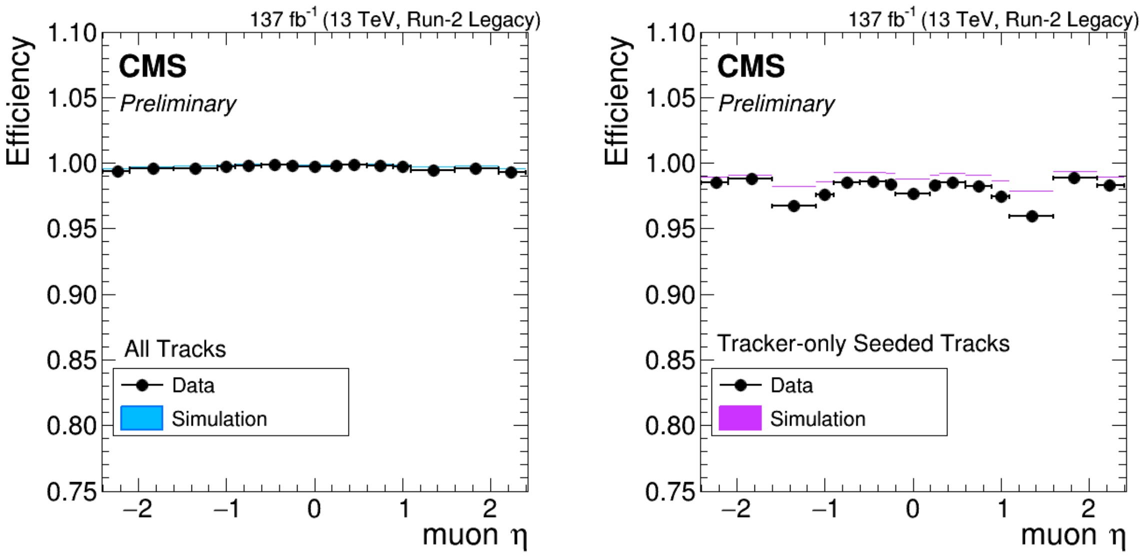


Figure 3.12: The muon tracking efficiency as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "tracker-only seeded tracks" (right). Run 2 Legacy data is represented by black dots and simulation by colored rectangles. The uncertainties shown are statistical.

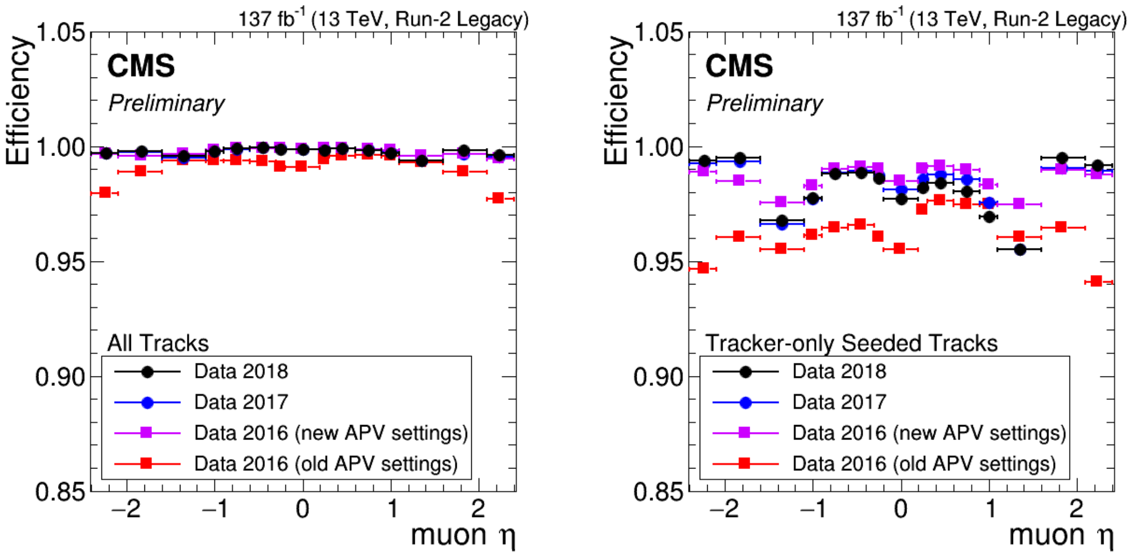


Figure 3.13: The comparison between the muon tracking efficiency for the different eras of Run 2 data as a function of pseudorapidity η of standalone muons using tracks in the "All-Tracks" collection (left) and "tracker-only seeded tracks" (right). The uncertainties shown are statistical.

Chapter 4

Effects of the electronic threshold on the performance of the RPC detector

As a robust handle to the signature of interesting events in the CMS, muons' trigger and reconstruction capabilities are essential. The muon system allows identifying muons produced in many standard model processes, such as top quark production, W and Z decay, Higgs boson studies, and searches beyond the Standard model. Therefore a robust and redundant spectrometer is needed to provide efficient muon reconstruction and identification. One of the detectors that play a critical role in triggering muons in the CMS experiment, both in the barrel and endcap regions, is the Resistive Plate Chambers. To optimize their performance, it is of primary importance to tune the electronic threshold of the front-end boards that read the signals from these detectors. This chapter introduces the results of a study that investigated the effects of variations in the electronic threshold voltage on the detector's RPC efficiency, cluster size, and intrinsic noise rate. The results presented in this chapter have been published in [107].

4.1 The Resistive plate chamber

Resistive Plate Chambers (RPC) [108] are gaseous parallel-plate detectors used as dedicated trigger detectors in the CMS in the barrel and endcap regions. The RPCs combine excellent time resolution with an acceptable spatial resolution, allowing the capability to tag an ionising event much shorter than the 25 ns between 2 consecutive LHC bunch crossings (BX); hence, it can assign the muon to the correct bunch crossing.

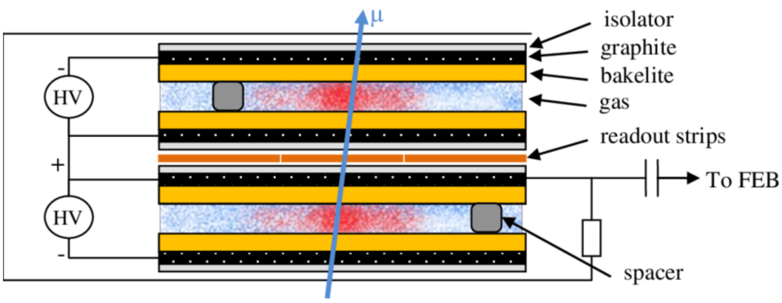


Figure 4.1: The layout of the double-gap RPC chamber. [109]

The CMS RPC chambers consist of two parallel High-Pressure Laminate (HPL) plates with a thickness of 2 mm and a resistivity of 1-6$\times$ 10^{10} Ω cm, separated by a 2 mm gas gap. A grid of polycarbonate spacers is distributed within the gap to provide rigidity and maintain the thickness of the gap. The HPL inner surfaces are covered with linseed oil, while the outer surfaces are coated with conductive graphite to make the high voltage and ground electrodes. The readout is performed via capacitive coupling of copper strips, separated from the graphite coating by an insulating PET film. The CMS RPC chambers are made of two gaps to form the so-called double-gap geometry, with pick-up strips in the middle. A layout of the double gap design is presented in Figure 4.1.

RPC operating principle is based on the ionization process like any other gas detector. When a muon passes through a gas gap, it ionizes the gas atoms, creating electron-ion pairs. The applied electric field at the gap causes the electrons to drift towards the anode. Because of their lower mass, the electrons have a higher drift velocity than the ions. They are accelerated toward the anode, ionizing more molecules in their path, creating an avalanche. The induced charges are collected on the readout strips, located between and shared by the two gas gaps. The readout electronics then measures the time of the signal with an intrinsic resolution of around 2 ns. Considering the uncertainty of the signal's propagation time along the strip, the RPC measures the time of the hit with an overall resolution of better than 3 ns. This time resolution is much smaller than the LHC bunch spacing of 25 ns and is precise enough to assign the muon track to the proper bunch crossing with high efficiency (above 95% during Run 2 [110]).

The CMS RPC operate with a humidified gas mixture composed of [72, 108, 111]: 95.2% C2H2F4 (Freon-R134a), 4.5% iC4H10 (isobutane) and 0.3% SF6 (Sulphur Hexafluoride). The electrons drift velocity in C2H2F4 promotes the excellent timing properties of RPCs. During the avalanche multiplication process, hard UV photons are produced due to electron-atom collisions during ion-electron recombination in the gas or on the cathode surface. Polyatomic molecules, such as iC4H10, are used as quencher gas since they limit the avalanche development by absorbing UV photons. SF6 is used in minimal quantities for its high electro-negativity. The compound absorbs any excess of electrons, and streamers are suppressed [112, 113]. Nevertheless, a fraction of SF6 higher than 1% would not bring any additional benefit in streamer reduction power but would result in a higher operating voltage.

The CMS RPC system consists of 1056 RPCs, organized into 4 stations covering a pseudorapidity range of $|\eta| < 1.2$ in the Barrel region (RB1 - RB4), and 4 stations for each endcap covering $0.9 < |\eta| < 1.9$ (RE1 - RE4) [72].

There are 480 RPC installed and evenly distributed on 5 wheels along the beamline, designated as wheel 0, wheel ±1, and wheel ±2, as shown in Fig. 4.2 (left). Each wheel consists of 4 muon stations (named RB1, RB2, RB3 and RB4) from inside to outside the CMS detector and is divided into 12 sectors (named S01-S12) in ϕ. The inner two stations consist of 2 layers of RPCs (RB1in, RB1out, RB2in and RB2out) with a DT in between, while the outer two stations consist of a single layer of RPCs (RB3 and RB4) and a layer of DTs.

The design of endcap RPCs is similar to that of barrel RPCs, except for the different shape and fine segmentation required for the forward regions with higher multiplicity [115]. A total of 576 RPCs

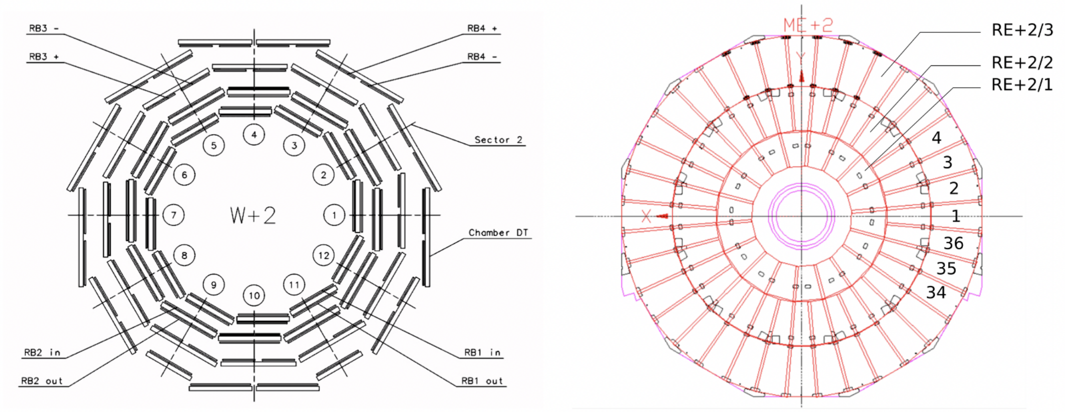


Figure 4.2: CMS RPC system layout in the Barrel for one Wheel W+2 (left) and in the endcap for one disk with 3 rolls (right). [114]

are installed in the endcap and distributed in 8 disks (4 disks per endcap, positive and negative), named RE±1, RE±2, RE±3 and RE±4, as shown in Fig. 4.2 (right). The two disks for the fourth station, both positive and negative, were installed during LS1 (2013-2014) [38]. Each disk is divided into 36 sectors along ϕ and has two concentric rings: ring 2 (R2) and ring 3 (R3). Each ring consists of 36 trapezoidal chambers spanning 10° and further divided into three rolls, denoted A, B, and C as η increases. By dividing the readout strips into three rolls, the TOP layer gap is divided into two gaps: Top Narrow (TN) and Top Wide (TW). The Bottom (BOT) layer consists of one gap instead.

4.2 Effect of changing front-end discrimination thresholds on RPC detector performance

Signals coming from each η partition of the RPC detector are read-out by means of a front-end electronic board, which allows only signals larger than a specific FE discriminator value to be transmitted to the trigger logic. Therefore, fine-tuning of these electronic thresholds affects some of the main detector characteristics, like efficiency, cluster size and detector intrinsic noise rate.

The CMS RPC system has been operating for several years, with electronic thresholds defined at the first commissioning. These are usually called "default thresholds", differ for each roll and front-end electronics board considered, and are stored in the configuration database. They will be henceforth indicated with $V_{thr-def}$. The values of $V_{thr-def}$ usually lie in the [205-240] mV range, where 1 mV corresponds to a signal of about 3.2 fC before pre-amplification [116].

In order to investigate the possibility to optimize the electronic thresholds, a dedicated scan has been performed during a data taking campaign with cosmic rays. In the scan, the electronic thresholds $V_{thr-app}$ actually applied to the front-end boards were changed, starting from ($V_{thr-def}$ - 5 mV) up to ($V_{thr-def}$ +15 mV), in five steps of 5 mV each. In this section, results about efficiency, cluster size, and detector intrinsic noise, obtained during this threshold scan, are reported.

4.2.1 Effect on RPC efficiency

RPC hit efficiency was measured using the segment extrapolation method [117]. The segments formed in the most adjacent DT (in the barrel) or CSC (in the endcap) chambers were extrapolated to the RPC chamber level. The selected segments are checked to belong to a standalone muon track reconstructed in the muon system, and the selected tracks have $p_T > 7$ GeV and quality defined by $\chi^2/\text{NDOF} < 8$. The extrapolated points are matched to the points reconstructed from the RPC hits in a range of four strips. The efficiency was calculated as the ratio between the number of matching and expected hits. The average barrel RPC efficiency measured during the threshold scan is summarized in Table. 4.1, for various values of $V_{thr-app}$. Due to rather small amount of data in the endcap chambers, a detailed study was possible in the barrel region only. As expected, efficiency decreases as $V_{thr-app}$ increases, of about 3.5% in the threshold range considered.

$V_{thr-app}$ - $V_{thr-def}$	-5 mV	0 mV	+5 mV	+10 mV	+15 mV
Average Efficiency	94.5 %	93.6 %	93.1 %	92.2 %	91.0 %

Table 4.1: RPC barrel average efficiency, for different applied electronic thresholds.

To exemplify this behaviour, RPC efficiencies of one roll in the barrel (left) and one in the endcap region (right), as a function of the difference between the applied and the default thresholds ($V_{thr-app}$ - $V_{thr-def}$), are shown in Fig. 4.3.

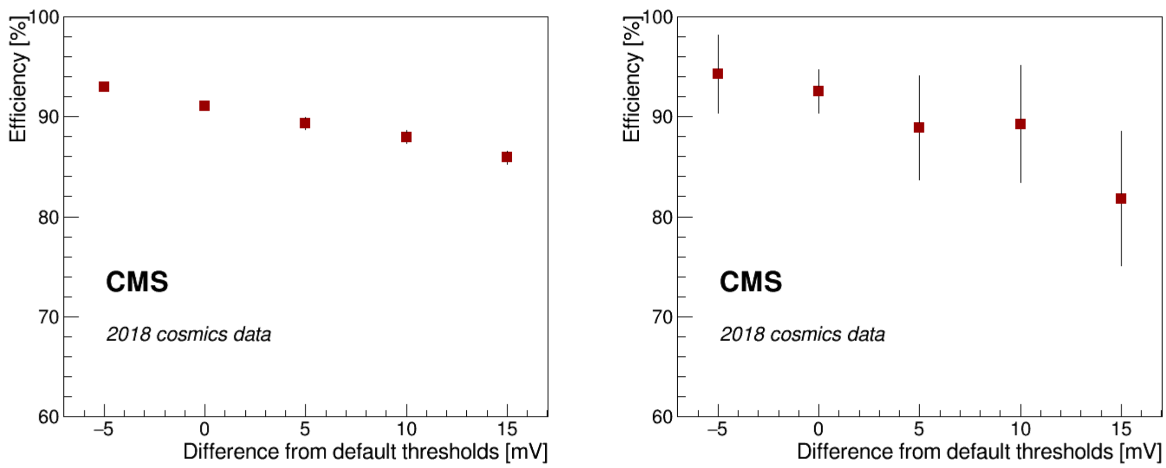


Figure 4.3: RPC efficiency as a function of the difference between $V_{thr-app}$ and $V_{thr-def}$. This figure refers to one roll for the barrel (left) and one in the endcap region (right). Error bars for the endcap roll are larger due to less amount of data.

Barrel RPC roll efficiency distributions, measured at the different $V_{thr-app}$ used during the voltage scan, are shown in Fig. 4.4 (left). Efficiency distributions present larger tails to the left of the peak as rolls' efficiency decrease with increasing the front-end discrimination thresholds. The RPC efficiency variation distribution between the efficiency values measured at $V_{thr-def}$, and the ones obtained with $V_{thr-app}$ = ($V_{thr-def}$ - 5 mV) is shown in Fig. 4.4 (right). Applying a threshold $V_{thr-app}$ 5 mV lower with respect to $V_{thr-def}$ results, in the barrel, in an average efficiency gain of about 0.9%. The tails in the distributions, containing few entries characterized by high efficiency variations, are caused by less amount of data.

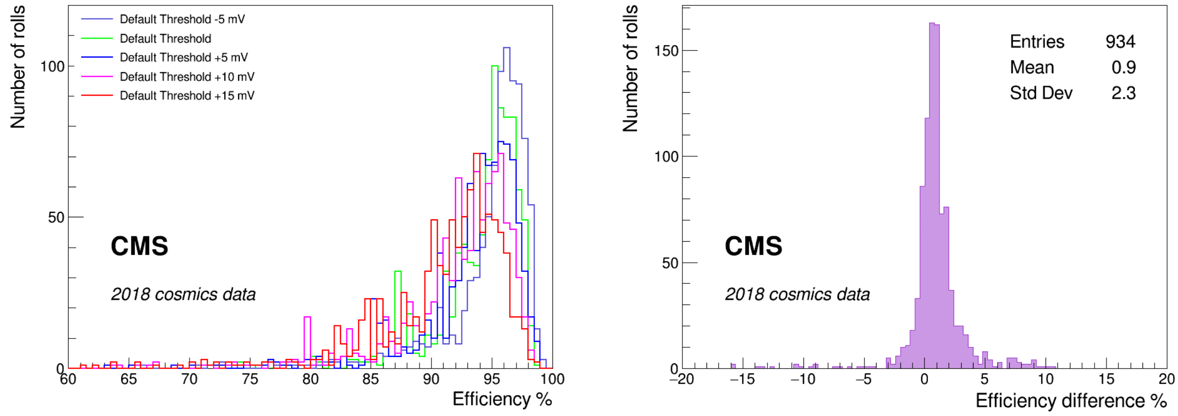


Figure 4.4: Left: Barrel RPC efficiency distributions, measured at the various $V_{thr-app}$ used during the threshold scan. Right: barrel RPC efficiency variation, measured at $V_{thr-app}$ = ($V_{thr-def}$ - 5 mV), with respect to the one obtained with $V_{thr-def}$.

The efficiency variation when decreasing the threshold voltage by 5 mV has also been studied for the different barrel regions of the RPC detector with respect to the beam pipe. Fig. 4.5 shows the efficiency variation when decreasing the threshold voltage by 5 mV, which results in an average efficiency gain of about 0.7% and 1.6% for the rolls from the innermost (closest to the beam pipe) RPC barrel layer RB1in (left), and for the outermost one - RB4 (right) respectively.

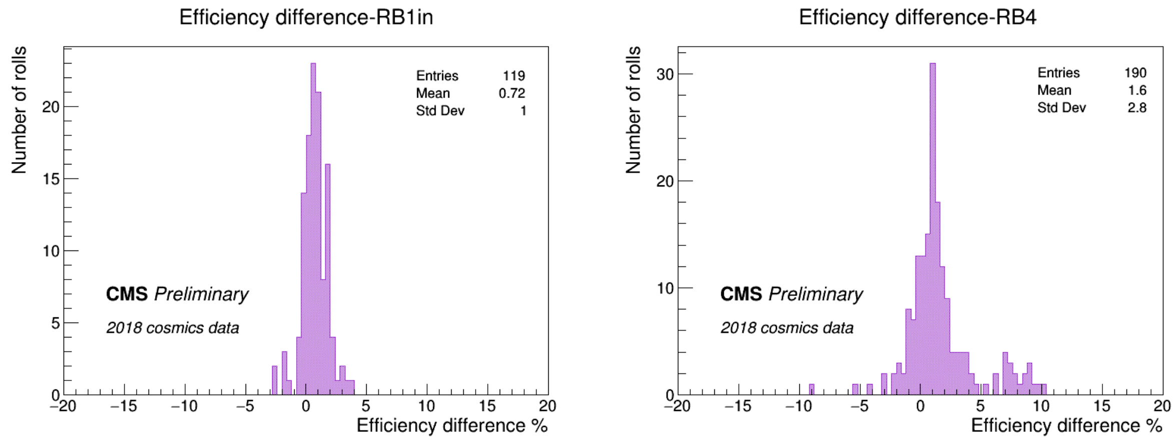


Figure 4.5: The Variation of the barrel RPC efficiency when decreasing the applied threshold voltage by 5 mV with respect to $V_{thr-def}$ for the rolls from the innermost (closest to the beam pipe) RPC barrel layer RB1in (left), and for the outermost one - RB4 (right).

4.2.2 Effect on RPC cluster size

One of the essential parameters for monitoring the RPC system's performance is the RPC cluster size [117], defined as multiple adjacent fired RPC strips that respond to the passage of charged particles. The coordinates of the RPC hits are determined at the center of the formed clusters of fired strips. A larger cluster size may affect the correct estimation of the deflection angle of the muon trajectory. To meet the requirements of the muon trigger, the cluster size of the RPC hits should not exceed 3 stripes. The mean values of the barrel RPC cluster size measured applying different values of $V_{thr-app}$ used during the threshold scan are summarized in Table. 4.2. As expected, cluster size decreases as $V_{thr-app}$ increases.

$V_{thr-app}$ - $V_{thr-def}$	-5 mV	0 mV	+5 mV	+10 mV	+15 mV
Average cluster size (# of strips)	1.76	1.70	1.67	1.62	1.58

Table 4.2: Average barrel RPC cluster size, for different values of the applied threshold $V_{thr-app}$

.

The RPC cluster size of one roll in the barrel and one in the endcap region, as a function of $V_{thr-app}$ is shown in Fig. 4.6. Again, this exemplifies what happens to the cluster size when increasing the electronic threshold in the whole system.

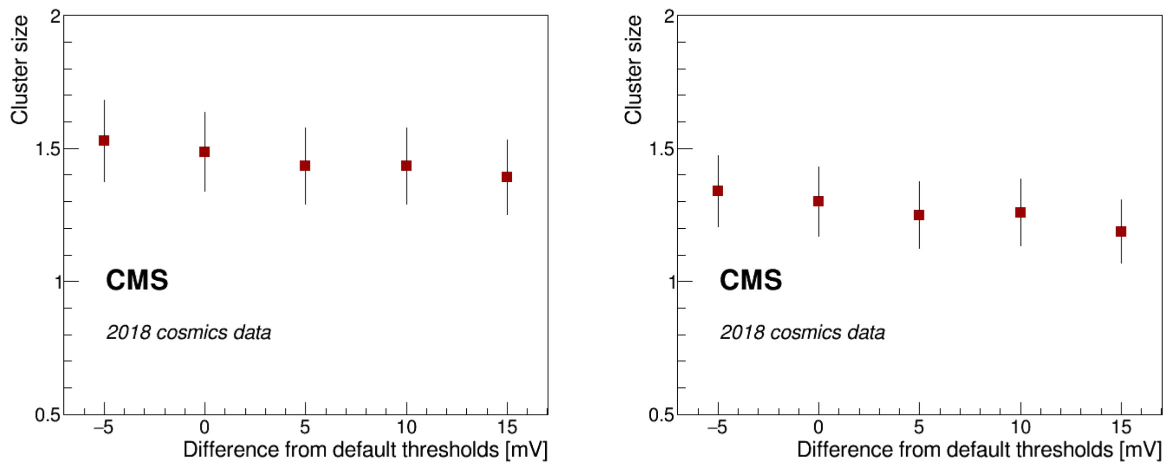


Figure 4.6: RPC Cluster size as a function of the difference between $V_{thr-app}$ and $V_{thr-def}$. Results for one barrel (left) and one endcap rolls (right) are shown.

The RPC cluster size distributions, measured at the different $V_{thr-app}$ used during the threshold scan, are shown in Fig. 4.7 (left). Applying a threshold discrimination lower by 5 mV with respect to $V_{thr-def}$ results in a slight increase of the cluster size, quantified in about 0.07 strips for the barrel. This is shown in Fig. 4.7 (right).

The cluster size variation when decreasing the threshold voltage by 5 mV has also been studied for the different barrel regions of the RPC detector with respect to the beam pipe. Fig. 4.8 shows the cluster size variation when decreasing the threshold voltage by 5 mV, which results in a slight cluster size increase of about 0.064 and 0.047 strips for the rolls from the innermost (closest to the beam pipe) RPC barrel layer RB1in (left) and for the outermost one - RB4 (right) respectively.

4.2.3 Effect on RPC intrinsic noise rate

The intrinsic noise rate of the RPC [118] is defined as the count rate normalized to the detector surface measured by the RPC chamber during cosmic data-taking. Under these conditions, the

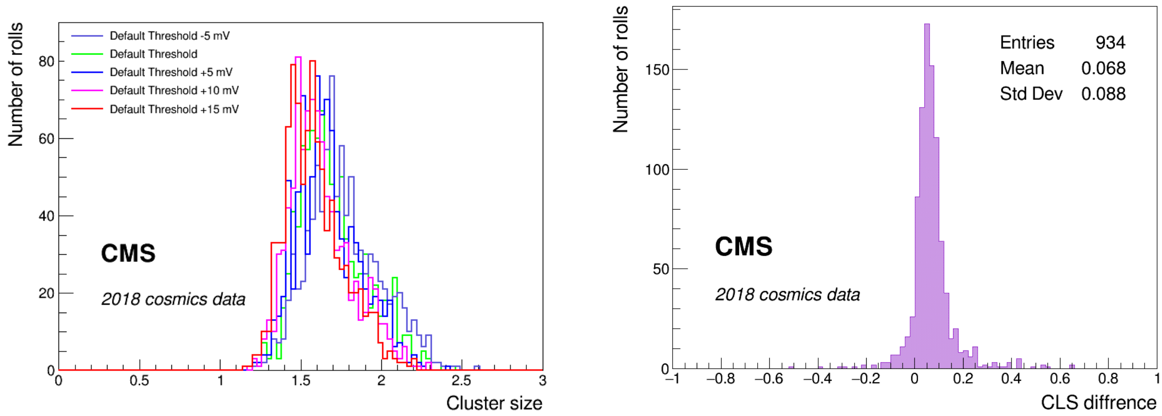


Figure 4.7: (Left) RPC cluster size distributions measured at the different applied thresholds $V_{thr-app}$ used during the threshold scan. (Right) Variation of the barrel RPC cluster size when decreasing the applied threshold voltage by 5 mV with respect to $V_{thr-def}$.

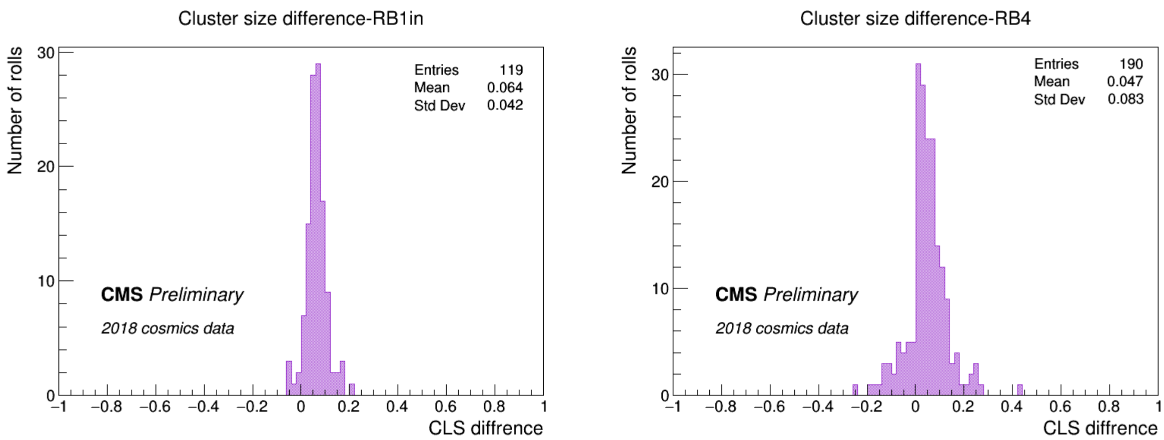


Figure 4.8: The Variation of the barrel RPC cluster size when decreasing the applied threshold voltage by 5 mV with respect to $V_{thr-def}$ for the rolls from the innermost (closest to the beam pipe) RPC barrel layer RB1in (left), and for the outermost one - RB4 (right).

RPC high voltage is at its working point, and cosmic muons are detected solely. Nevertheless, other contributions to the intrinsic rate of the detector may be detected, caused by the noise of the high voltage system, tensions in the cables, etc. Therefore, it is essential to keep this noise as low as possible, with a maximum of 5 Hz/cm^2 required by CMS. The intrinsic RPC noise rate as a function of (V$_{thr-app}$- V$_{thr-def}$) is shown in Fig. 4.9 for one roll in the barrel and one in the endcap region. A quadratic polynomial function is used for the fit.

The distribution of the difference of the intrinsic noise rate measured for the RPC chambers in the barrel when applying an electronic threshold 5 mV lower with respect to V$_{thr-def}$ is shown in Fig. 4.10. On average, applying 5 mV lower thresholds with respect to V$_{thr-def}$ results in an intrinsic noise rate increase of about 0.04 Hz/cm^2, which represents around 10% increase with noise rate measured at the default threshold.

The results presented in this chapter and obtained during the 2018 threshold scan of the CMS RPC system are in agreement with the expectations: a decrease in the applied discrimination thresholds

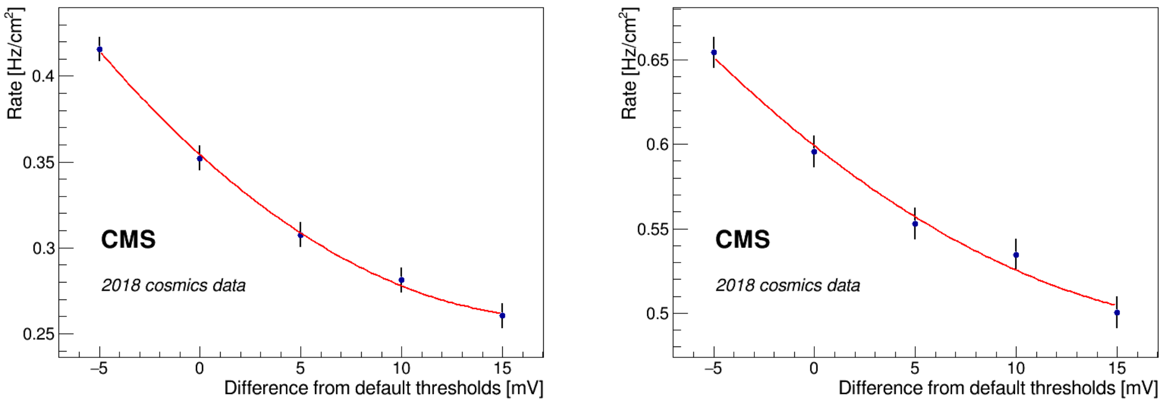


Figure 4.9: RPC intrinsic noise rate as a function of the difference between $V_{thr-app}$ and $V_{thr-def}$. One roll for the barrel (left) and one for the endcap region (right) are shown. In both cases, a quadratic polynomial function is used for the fit.

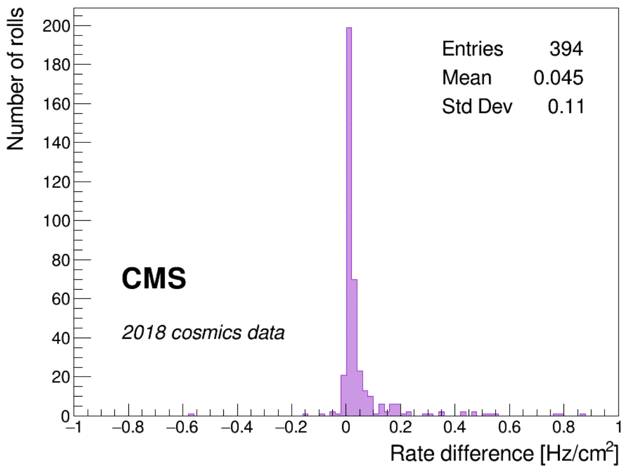


Figure 4.10: Variation of the intrinsic barrel RPC rate, measured when applying $V_{thr-def}$ and ($V_{thr-def}$ -5 mV), which results in an intrinsic noise rate increase of about 0.04 Hz/cm^2.

increases the RPC efficiency, which is a benefit for the system but, at the same time, induces an increase in cluster size and intrinsic noise, which could deteriorate spatial resolution and accidentals, respectively. The opposite happens when the threshold is increased. Nevertheless, the $\approx 0.9\%$ gain in efficiency consequent to a 5 mV decrease of the threshold with respect to the default one is to be preferred w.r.t. the corresponding minimal increase in cluster size, of around 0.07 strips, and intrinsic noise, roughly 0.04 Hz/cm^2. This would induce to consider thoughtfully adopting such a solution when the CMS RPC system is restarted after the current shutdown.

Chapter 5

Baseline selections for VBF analysis

The present analysis is based on the reconstruction of the SM Higgs boson produced through the VBF production channel, and decay to $ZZ \rightarrow 4l$. The typical physics objects, which are the outcome of this analysis, are electrons, muons, jets, as well as photons used for final-state radiation (FSR) recovery. The reconstruction of these physics objects are described in Section 2.4. This chapter describes the baseline selections used for the VBF analysis. The list of data sets used, the simulated signal and background samples are described. It also reports on the selection criteria applied to identify and isolate the physics objects used in this analysis.

5.1 Data Samples and Triggers

This analysis is based on proton-proton collision data collected by the CMS detector at the LHC during Run 2 at $\sqrt{s}$ = 13 TeV, corresponding to an integrated luminosity of 137 fb^{-1}. Three different data-taking periods were used for the data collection: 2016, 2017, and 2018 corresponding to an integrated luminosity of 35.9, 41.5, and 59.7 fb^{-1}, respectively [119–121].

The collision events are selected by high-level trigger algorithms that require the presence of leptons that satisfy loose identification and isolation constraints. The primary triggers select either a pair of electrons or muons or an electron and a muon. The minimal transverse momentum of the leading and sub-leading leptons has changed over the years, taking into account the different conditions for data acquisition and is summarized in Table 5.1. To maximize the coverage of the $H \rightarrow ZZ \rightarrow 4l$ phase space, triggers that require three leptons with relaxed transverse momentum thresholds and no isolation requirement, as well as isolated single-electron and single-muon triggers, are used.

	e/e (GeV)	μ/μ (GeV)	e/μ, μ/e (GeV)
2016	17/12	17/8	17/8, 23/8
2017	23/12	17/8	23/8, 23/12
2018	23/12	17/8	23/8, 23/12

Table 5.1: The minimal p_T of the leading/subleading leptons used in the main high-level trigger algorithms; the di-electron (e/e), dimuon (μ/μ), and electron-muon (e/μ, μ/e) used in the $H \rightarrow ZZ \rightarrow 4l$ analysis for 2016, 2017, and 2018 datasets.

5.1.1 Trigger Efficiency

The efficiency of the trigger used in this analysis is measured from data using a method based on the "tag and probe" technique, described in Section. 3.3, by considering 4ℓ events triggered by single-lepton triggers. One of the four reconstructed leptons (the "tag") is geometrically matched to a trigger object passing the final filter of one of the single-muon or single-electron triggers. The other three leptons are used as "probes". For each 4ℓ event, there are up to 4 possible tag-probe combinations, and all possible combinations are counted in the denominator of the efficiency. For each of the three probe leptons, all matching trigger filter objects are collected. Then, the matched trigger filter objects of the three probe leptons are combined to reconstruct each of the triggers used in the analysis. If any of the analysis triggers can be formed using the probe leptons, the set of probes is also included in the numerator of the efficiency.

The efficiency plotted as a function of the minimum p_{T} of the three probe leptons in the data and MC using this method can be seen in Fig. 5.1, which shows the trigger efficiency in the 2016 data for different final states. The MC efficiency describes the data well within the statistical uncertainties.

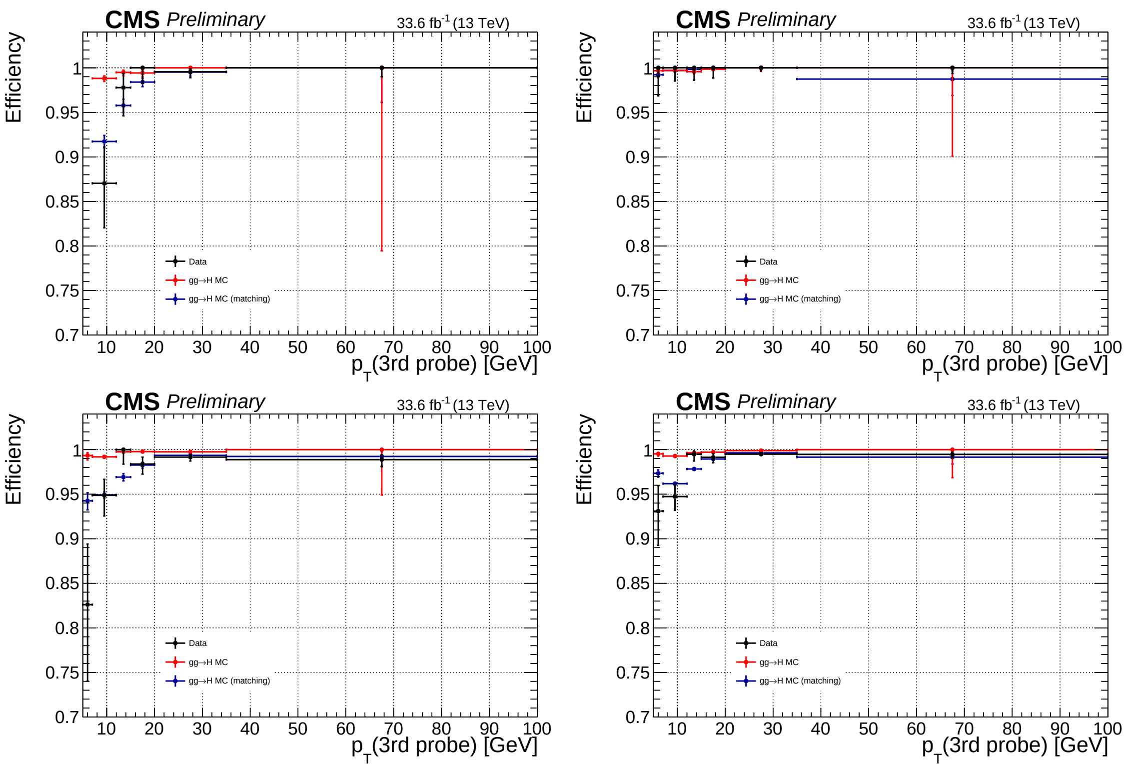


Figure 5.1: Trigger efficiency measured in 2016 data using 4ℓ events collected by single lepton triggers for the $4e$ (top left), 4μ (top right), $2e2\mu$ (bottom left) and 4ℓ (bottom right) final states [122].

5.2 Simulated samples

MC Simulation of signal (i.e., the final states corresponding to the reaction under investigation) and relevant background processes (i.e., all categories of events that are not due to the reaction under investigation but have final states similar to those sought) is essential to any physics analysis. It can be used to evaluate signal shape, optimise event selection, estimate background rates, estimate expected yields, and evaluate the acceptance and systematic uncertainties.

In this analysis, the Higgs boson produced by VBF is considered a signal, while the other production modes for the Higgs boson are considered backgrounds. The SM Higgs samples are simulated at NLO in perturbative QCD (pQCD) using the POWHEG V2 generator [123–125] for the five main production modes: gluon fusion (ggH), including quark mass effects [126], vector boson fusion (VBF) [127], associated production with a vector boson (WH, ZH) [128], and associated production with a pair of top quarks ($t\bar{t}$H) [129]. In addition to the five main production processes, the contribution of the associated production with a pair of bottom quarks ($b\bar{b}$H) is simulated at LO with JHUGEN. For all the samples, the decay of the Higgs boson into ZZ and finally into four leptons is described using the JHUGEN generator [130]. The theoretical predictions used for the various production and decay modes are summarized in Ref. [131]. The parton shower and the hadronization are carried out by PYTHIA. The list of SM Higgs processes and the corresponding cross-sections times branching ratio are shown in Tab. 5.2.

The ZZ background contribution from $q\bar{q}$ annihilation is generated at NLO pQCD using POWHEG V2 [132], with the same settings as for the Higgs signal. As this simulation covers a large range of ZZ invariant masses, dynamical QCD factorization and renormalization scales have been chosen to be equal to m_{ZZ}. The $gg \rightarrow ZZ$ process is simulated at LO with MCFM [133, 134].

In addition, the WZ, triboson backgrounds ZZZ, WZZ and WWZ, and $t\bar{t}$ + vector bosons (V) are generated at NLO using MadGraph5 (aMC@NLO), while the smaller $t\bar{t}$Z, $t\bar{t}$WW and $t\bar{t}$ZZ background processes are simulated at LO using MadGraph5 (aMC@NLO). The events containing Z bosons with associated jets (Z+jets), also called Drell-Yan (DY), and the $t\bar{t}$ background are simulated at NLO using MadGraph5 (aMC@NLO).

All signal and background event generators are interfaced with PYTHIA [96] to simulate the multi-parton shower and hadronization effects. The generated events are processed through GEANT4 [99, 135] to simulate the CMS detector in detail and then reconstructed with the same algorithms used for the data.

5.2.1 Modelling the 3^{rd} jet in VBF-HJJJ sample

The third jet in the VBF sample (VBF-H2J), mentioned above, is generated at LO. Currently, a better description for VBF Higgs production associated with three hard jets at NLO (VBF-H3J) has been provided [136] based on the matrix elements of Ref.[137] extracted from the VBF NLO code [138], [139]. These implementations rely on the VBF approximation.

The Feynman diagrams relevant for this process are shown in Fig. 5.2: considering the subprocesses $qQ \rightarrow qQgH$ and crossed subprocesses with the vector-boson exchange in the t-channel (time-channel). We explicitly exclude weak boson exchanges in the s-channel (space channel) and therefore leave aside higgsstrahlung processes, i.e., the production of V Hj followed by decay $V \rightarrow jj$. We also neglect any interference of VBF and higgsstrahlung (in the case of identical fermion flavours)

Process	generator + PS	$\sigma \cdot BR$
	SM Higgs MC Samples	
$gg \to H \to ZZ \to 4l$	Powheg + Pythia	12.18 fb
$q\bar{q} \to Hq\bar{q} \to ZZq\bar{q} \to 4lq\bar{q}$	Powheg + Pythia	1.044 fb
$q\bar{q} \to W^+H \to W^+ZZ \to 4l + X$	Powheg + Pythia	0.232 fb
$q\bar{q} \to W^-H \to W^-ZZ \to 4l + X$	Powheg + Pythia	0.147 fb
$q\bar{q} \to ZH \to ZZZ \to 4l + X$	Powheg + Pythia	0.668 fb
$gg \to t\bar{t}H \to t\bar{t}ZZ \to 4l + X$	Powheg + Pythia	0.393 fb
$gg \to b\bar{b}H \to b\bar{b}ZZ \to 4l + X$	JHUGEN + Pythia	0.135 fb
	SM Backgrounds MC Samples	
$q\bar{q} \to ZZ \to 4l$	Powheg + Pythia	1.256 pb
$q\bar{q} \to ZZ \to 4l + jets(EWK)$	MadGraph + Pythia	0.0004404 pb
$gg \to ZZ \to 4e$	MCFM + Pythia	0.00159 pb
$gg \to ZZ \to 4\mu$	MCFM + Pythia	0.00159 pb
$gg \to ZZ \to 4\tau$	MCFM + Pythia	0.00159 pb
$gg \to ZZ \to 2e2\mu$	MCFM + Pythia	0.00319 pb
$gg \to ZZ \to 2e2\tau$	MCFM + Pythia	0.00319 pb
$gg \to ZZ \to 2\mu2\tau$	MCFM + Pythia	0.00319 pb
$Z \to ll + jets$	MadGraph5 (aMC@NLO) + Pythia	18610 pb
	MadGraph5 (aMC@NLO) + Pythia	6104 pb
$t\bar{t}$	MadGraph5 (aMC@NLO) + Pythia	815.96 pb
$WZ \to 3l\nu$	Powheg + Pythia	4.430 pb
ZZZ	MadGraph5 (aMC@NLO) + Pythia	0.01398 pb
WWZ	MadGraph5 (aMC@NLO) + Pythia	0.1651 pb
WZZ	MadGraph5 (aMC@NLO) + Pythia	0.05565 pb
$t\bar{t}$ + ZZ	MadGraph + pythia	0.001572 pb
$t\bar{t}$+WW	MadGraph + pythia	0.007883 pb
$t\bar{t}$+Z	MadGraph + pythia	0.259 pb

Table 5.2: MC simulated samples and their respective cross section.

since these interference effects are minimal in the phase space region relevant for VBF observation at the LHC [140, 141].

A private VBF-H3J sample has been produced using the Powheg-Box V2 MC generator, and the events were processed through the complete CMS simulation chain adopted for each year of Run 2. The cross-section times branching ratio for the generated sample is of 0.16 fb.

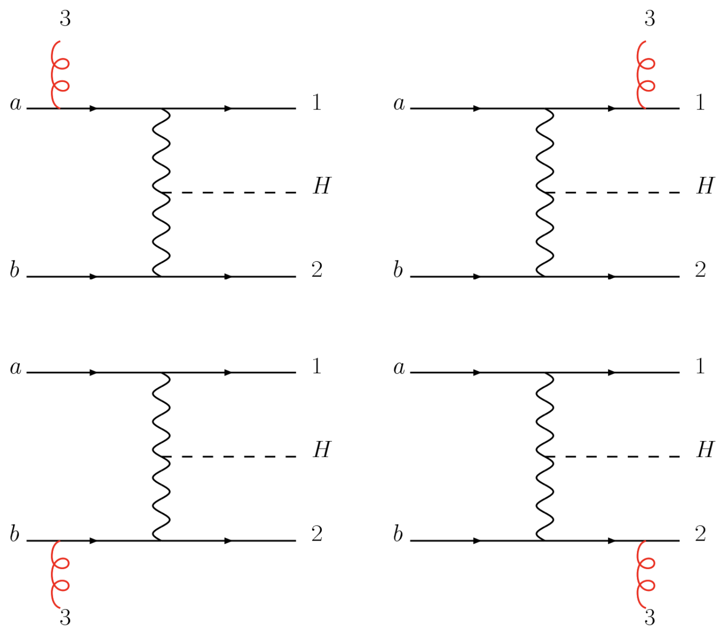


Figure 5.2: Lowest-order Feynman diagrams for $pp \to Hjjj$ via VBF Higgs production. [142]

5.2.2 MC Event Reweighting

The simulated samples include other particles from multiple additional interactions recorded together with the interaction of interest in the same and neighboring bunch crossings, the so-called pileup. The pileup distribution in MC samples does not precisely match the pileup distribution during data taking, which depends on experimental parameters such as the beam conditions and the instantaneous luminosity. For each year, corresponding simulation samples are reweighted to match the pileup distribution in data. An example of reweighting procedure for 2018 data is shown in Fig. 5.3.

In addition to the pileup weight, each event in the MC samples must be assigned a weight, which relies on the process production cross-section (σ), the integrated luminosity of data (L), and the total number of events generated in the MC sample. The MC event is weighted by a factor (w) given by

$$w = \frac{\sigma \cdot L}{N} \tag{5.1}$$

5.3 Physics Objects Selections

The present analysis is based on the reconstruction of the SM Higgs boson in the VBF production channel and the decay chain $H \to ZZ \to 4l$. Therefore, the final state should consist of two pairs

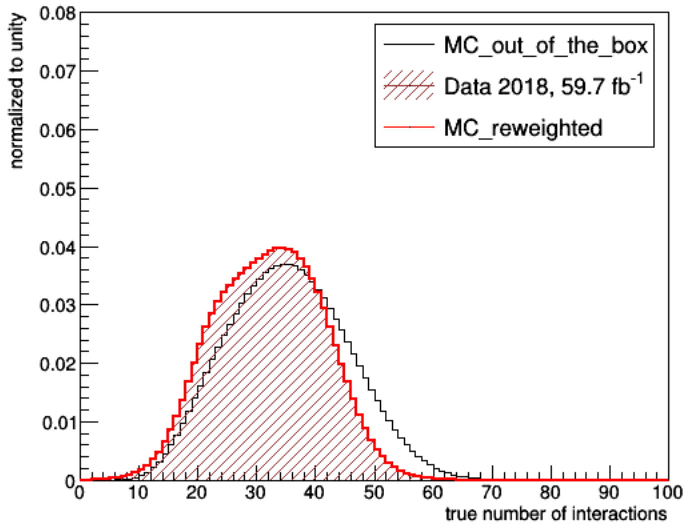


Figure 5.3: Distribution of pileup in 2018 Data and MC, before and after applying PU weights in comparison with data.

of opposite-charged isolated leptons and at least two jets. This decay channel requires sensitivity to a low mass Higgs, for which at least one of the leptons has a low p_T; hence it requires very efficient lepton reconstruction and identification. In the following sections, the selection criteria applied to the physics objects: electrons, muons, and jets, as well as photons used for FSR recovery, are presented.

5.3.1 Electrons

After the electron reconstruction process described in Section 2.4.3, we obtain a set of electron candidates, which could be real electrons or fake objects that pass the electron reconstruction process. The efficiency of electron reconstruction [143] is very high, which means that we lose only a small fraction of real electrons, but there could be some fake electrons. Therefore, a set of selection criteria is applied to the reconstructed electrons to reduce the number of fake objects while preserving the real electrons as much as possible. The criteria applied to the electrons are described below.

5.3.1.1 Electron kinematics and impact parameter cuts

The electrons are required to be within the detector acceptance ($|\eta| < 2.5$) and have a transverse momentum $p_T > 7$ GeV to reduce the QCD background. Loose constraints are then used on the impact parameter, which is the closest approach distance between the measured track of the electron and the primary vertex of the event, to identify electrons originating from the primary vertex or a secondary vertex:

$$|d_{xy}| < 0.5 \text{ cm and } |d_z| < 1 \text{ cm}.$$

where d_{xy} and d_z are the electron transverse and longitudinal impact parameters with respect to the primary vertex, respectively. The electrons passing the mentioned cuts are referred to as *loose electrons*.

To ensure that the leptons are consistent with a common primary vertex (PV), explained in Section. 6.1.2, we require that they have an associated track with a small impact parameter with respect to the primary vertex of the event. Therefore, the significance of the impact parameter to the event vertex is required to satisfy,

$$|\mathrm{SIP}_{3\mathrm{D}} = \frac{\mathrm{IP}}{\sigma_{\mathrm{IP}}}| < 4$$

where IP is the lepton impact parameter in three dimensions at the point of closest approach with respect to the primary interaction vertex, and σ_{IP} is the associated uncertainty. By applying these cuts, we can suppress the background electrons originating from the displaced vertices. Electrons satisfying this cut are defined as *primary electrons*.

5.3.1.2 Electron Identification and Isolation

The reconstructed electrons are identified and isolated using a distributed gradient boosting library optimized for eXtreme Gradient Boosting (XGBoost) [144], designed to be highly efficient, flexible, and portable. It implements machine learning algorithms in the Gradient Boosting framework and exploits observables from the electromagnetic cluster, the matching between the cluster and the electron track, observables based on tracking measurements, as well as particle-flow (PF) isolation sums. The list of observables used in training is summarized in Table 5.3.

Feeding the XGBoost with this wide range of variables allows it to distinguish between a good electron and hadronic jets that can mimic the electron signature in the detector and reduce the number of fake electrons, especially in the low p_T region without a large impact on the signal efficiency.

The classifier is trained on a DY with jets MC sample, where reconstructed electrons with a geometrical match with generated electrons originating from Z represent signal (real electrons), and the reconstructed electrons that don't match generated electrons originating from the Z are considered as backgrounds (fake electrons). The signal electrons from the Higgs and DY samples have the same kinematics since the electrons come from Z decay in both samples. Fig. 5.4 shows an example for the output of the multiclassifier discriminant, i.e., MVA score for prompt electrons from DY events and misidentified electrons originating from jets in DY events.

The training has been done in different bins of electron's p_T to improve the identification performance and different bins of electron's η since the electron reconstruction is done separately for barrel and endcap regions. Tables 5.4, 5.5, and 5.6 summarize the cuts values applied to the MVA output for 2016, 2017, 2018 training, respectively. The loose electrons used in the VBF Higgs analysis have to pass this MVA identification and isolation working point.

5.3.1.3 Electron Efficiency

The efficiency of the cuts could behave differently in data and simulation, and therefore these differences in efficiency need to be investigated and quantified. Reliable measurement of the efficiency of electron reconstruction, identification, isolation, and impact parameter cuts with both data and

observable type	observable description
cluster shape	• RMS of the energy-crystal number spectrum along η and φ; $\sigma_{i\eta i\eta}$, $\sigma_{i\varphi i\varphi}$ • Super cluster width along η and ϕ • Ratio of the hadronic energy behind the electron supercluster to the supercluster energy, H/E • Sum of the seed and adjacent crystal over the super cluster energy R_9 • Energy fraction in pre-shower (used for endcap training bins:) E_{PS}/E_{raw} • Circularity $(E_{5\times5} - E_{5\times1})/E_{5\times5}$, where ($E_{5\times5}$ represents the energy computed in the 5 × 5 block of crystals centered on the highest energy crystal of the cluster seed, and $E_{5\times1}$) is the energy calculated in the strip of crystals containing it.
track-cluster matching	• Ratio between the energy of the ECAL cluster closest to the electron track extrapolated to ECAL and the track momentum at the outermost track position E_{ele}/p_{out}. • Ratio between the supercluster energy and the track momentum at the innermost track position E_{tot}/p_{in}. • Deviation of the supercluster energy from the electron momentum obtained by combining ECAL and tracker information $1/E_{tot} - 1/p_{in}$. • Pseudorapidity (Azimuthal) distance between the energy weighted position of the supercluster and the track position extrapolated from the innermost track position $\Delta\eta_{in}$ ($\Delta\varphi_{in}$) • Pseudorapidity distance between the seed cluster position and the electron track extrapolation to the ECAL $\Delta\eta_{seed}$.
tracking	• Fractional momentum loss $f_{brem} = 1 - p_{out}/p_{in}$ • Reduced χ^2 of the KF and GSF track χ^2_{KF}, χ^2_{GSF} • Number of hits of the KF and GSF track N_{KF}, N_{GSF} • Probability transform of conversion vertex fit χ^2 • Number of expected but missing inner hits
isolation	• Particle-flow photon isolation sum • Particle-flow neutral hadrons isolation sum • Particle-flow charged hadrons isolation sum
For PU-resilience	• Mean energy density in the event: ρ

Table 5.3: Overview of input variables to the identification classifier.

2016 Datasets			
minimum BDT score	$\lvert\eta\rvert < 0.8$	$0.8 < \lvert\eta\rvert < 1.479$	$\lvert\eta\rvert > 1.479$
$5 < p_T < 10$ GeV	0.9503	0.9461	0.9387
$p_T > 10$ GeV	0.3782	0.3587	-0.5745

Table 5.4: Minimum BDT score required for passing the electron identification used for VBF Higgs analysis, for 2016 samples.

2017 Datasets			
minimum BDT score	$\lvert\eta\rvert < 0.8$	$0.8 < \lvert\eta\rvert < 1.479$	$\lvert\eta\rvert > 1.479$
$5 < p_T < 10$ GeV	0.8521	0.8268	0.8694
$p_T > 10$ GeV	0.9825	0.9692	0.7935

Table 5.5: Minimum BDT score required for passing the electron identification used for VBF Higgs analysis, for 2017 samples.

2018 Datasets			
minimum BDT score	$\lvert\eta\rvert < 0.8$	$0.8 < \lvert\eta\rvert < 1.479$	$\lvert\eta\rvert > 1.479$
$5 < p_T < 10$ GeV	0.8956	0.9111	0.9401
$p_T > 10$ GeV	0.0424	0.0047	-0.6042

Table 5.6: Minimum BDT score required for passing the electron identification used for VBF Higgs analysis, for 2018 samples.

MC is obtained using the tag and probe technique, explained in Section 3.3. These measurements are then used to build data/MC scale factors to correct the simulation.

The electron efficiency is measured in two steps: firstly, the electron reconstruction efficiency ϵ_{reco} is the GSF tracking efficiency calculated as the ratio of superclusters reconstructed as electrons

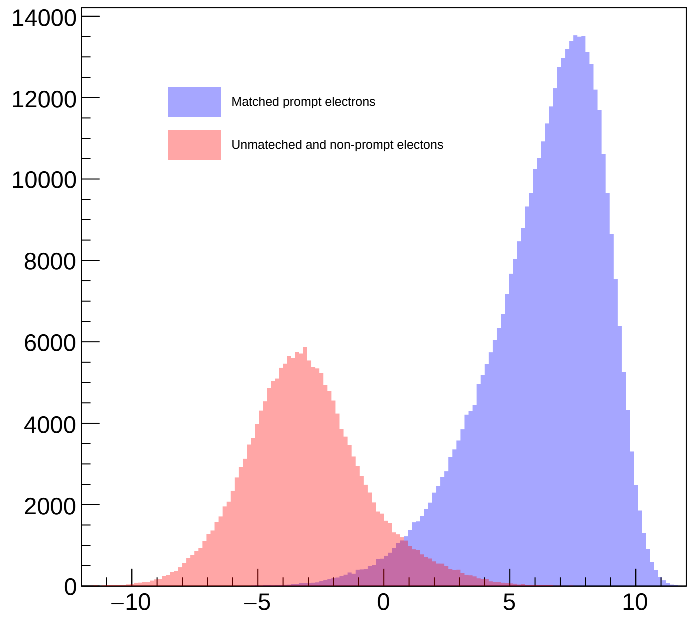


Figure 5.4: The output of the multiclassifier discriminant for prompt electrons matched to truth electrons from Z decay (blue) and for misidentified electrons (red). Drell-Yan with jets MC sample is used in training. [122]

(passing probes) to the superclusters (all probes). Then, the electron selection efficiency $\epsilon_{sel|reco}$ is measured as the selection cut efficiency on the electrons that have already passed the reconstruction step. The electron selection efficiency is measured as a function of the electron p_T and its η^{SC}, separately for gap electrons and non-gap electrons.

Fig. 5.5 shows an example of the electron reconstruction efficiency in 2017 data as a function of electron p_T and Fig. 5.6 shows the electron selection efficiency as a function of electron p_T for non-gap electrons (left) and gap electrons (right), together with the corresponding data/MC ratio (bottom).

5.3.2 Muons

Following the muon reconstruction process described in Section 2.4.2, we obtain a set of muon candidates, which could be real muons or fake objects that pass the muon reconstruction process. The efficiency of muon reconstruction [73] is very high, which means that we lose only a tiny fraction of real muons, but there could be some fake muons. Therefore, a set of selection criteria is applied to the reconstructed muons to reduce the number of fake objects while preserving the real muons as much as possible. The criteria used for the muons are described below.

5.3.2.1 Muon kinematics and impact parameter cuts

As for the electrons, all muons should be in the detector acceptance $|\eta| < 2.4$ and have $p_T > 5$ GeV. The p_T cut for muons is slightly lower than for electrons because the reconstruction of muons with low p_T is somewhat more efficient and thus lower momentum values can be measured. To

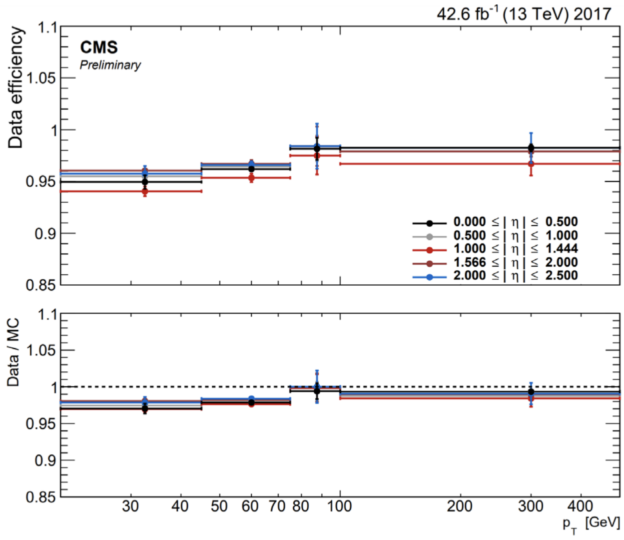


Figure 5.5: Electron reconstruction efficiency as a function of electron p_T in 2017 data (top) and data/MC ratio (bottom). [145]

distinguish between real and fake muons (from cosmic rays or from in-flight hadron decays), the muons should satisfy the impact parameters constraints:

$$|d_{xy}| < 0.5 \text{ cm} \ \& \ |d_z| < 1 \text{ cm} \ \& \ |\text{SIP}_{3D}| < 4$$

where d_{xy} and d_z are the muon transverse and longitudinal impact parameters with respect to the primary vertex, respectively, and SIP_{3D} is the significance of the impact parameter. Muon passing the kinematic and impact parameter cuts are marked as *loose muons*.

5.3.2.2 Muon Identification

Muons have to be reconstructed using either the Global Muon or the Tracker Muon algorithm, as described in Section 2.4.2. Standalone Muon tracks that are reconstructed only in the muon system are rejected.

Loose muons with p_T below 200 GeV are considered as identified muons if they also pass the PF muon ID described in Section 2.4.1. Loose muons with p_T above 200 GeV are considered identified muons if they pass the PF muon ID or the Tracker High-p_T ID whose definition is given in Table 5.7.

This relaxed definition is used to increase signal efficiency when searching for high masses. When a very heavy resonance decays to two Z bosons, both bosons will be very boosted. In the lab frame, the leptons coming from the decay of a highly boosted Z will be nearly collinear, and the PF ID loses efficiency for muons separated by about $\Delta R < 0.4$, which roughly corresponds to muons originating from Z bosons with $p_T > 500$ GeV.

An additional "ghost-cleaning" step is done in order to avoid the situation that a single muon incorrectly is being reconstructed as two or more muons:

- Tracker Muons that are not Global Muons are required to be "arbitrated", i.e., assigned to at least two segments in the outer muon detectors;

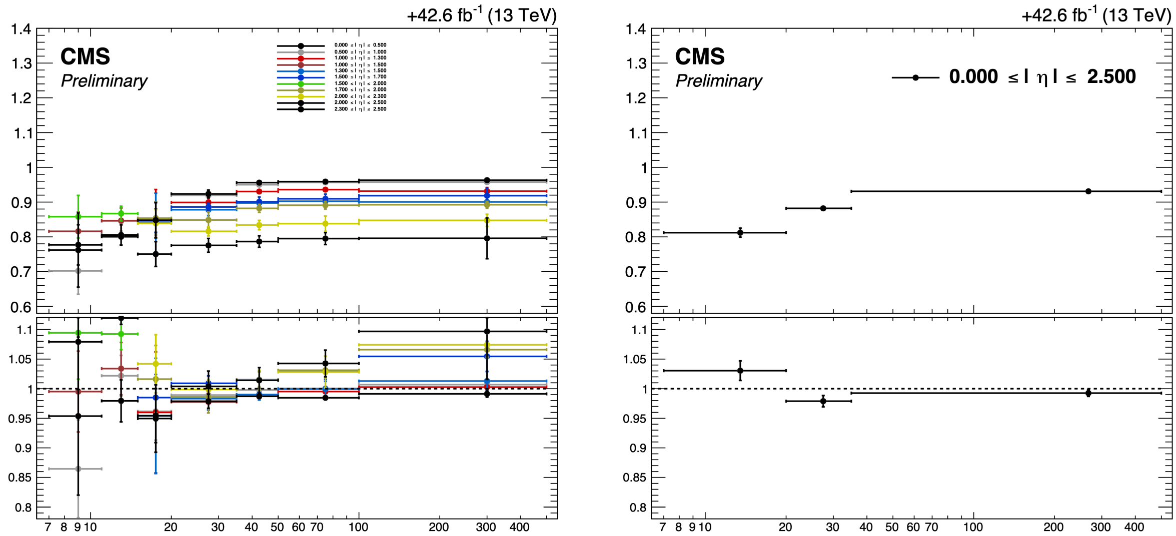


Figure 5.6: The electron selection efficiency as a function of electron p_T for non-gap electrons (left) and gap electrons (right), and the corresponding data/MC ratio (bottom). [122]

Requirements	Description
Good p_T measurement	$p_T/\sigma_{p_T} < 0.3$
Muon station matching	Muon is matched to segments in at least two muon stations
Vertex compatibility (x-y)	$d_{xy} < 2mm$
Vertex compatibility (z)	$d_z < 5mm$
Tracker hits	Hits in at least six tracker layers
Pixel hits	At least one pixel hit

Table 5.7: The requirements for a muon to pass the tracker high-p_T ID.

- If two muons share 50% or more of their segments, the muon with the lower quality is removed.

5.3.2.3 Muon Isolation

For the muons, a PF-based isolation is applied. The so-called $\Delta\beta$ correction is used to subtract the pileup contribution for the muons by using the $\Delta\beta$ correction, where

$$\Delta\beta = \frac{1}{2} \cdot \sum_{PU}^{charged\ had.} p_T \tag{5.2}$$

gives an estimate of the energy deposit of neutral particles (hadrons and photons) from pileup vertices. The relative isolation for muons is then defined as:

$$\text{RelPFiso} = \frac{\sum^{charged\ had.} p_T + max(\sum^{neutral\ had.} E_T + \sum^{photon} E_T - \Delta\beta, 0)}{p_T^{lepton}} \tag{5.3}$$

The operating point for the isolation of muons was set as RelPFiso $(\Delta R = 0.3) < 0.35$.

5.3.2.4 Muon Efficiency

The muon efficiencies are measured using the Tag and Probe (T&P) method (see Section 1) on $Z \to \mu^+\mu^-$ and $J/\Psi \to \mu^+\mu^-$ events in bins of p_T and η. The Z sample is used to measure the efficiency of muon reconstruction and identification at high p_T and the efficiency of isolation and impact parameter requirements at all p_T. The J/Ψ sample is used to measure the reconstruction efficiency at low p_T, as it benefits from better purity in this kinematic range.

The muon tracking efficiency ϵ_{trk} was evaluated using $Z \to \mu^+\mu^-$ events. The probe is defined as tracks reconstructed in the muon system "standalone muon", and the passing probes are the standalone muons that match a track that fulfills minimum quality requirements. For $p_T >$ 10 GeV, the "all-tracks" are used for matching, while for $p_T <$ 10 GeV, the "tracker-only seeded tracks" are used. Full details of this measurement and the results are described in Chapter 3.

The reconstruction and identification efficiency $\epsilon_{id|trk}$ results for muons with $p_T >$ 20 GeV were derived from $Z \to \mu^+\mu^-$ events. The probe in this measurement are tracks reconstructed in the inner tracker, and the passing probes are the tracks reconstructed as global or tracker muon (defined in Section 2.4.2). The efficiency for low p_T muons was derived from $J/\Psi \to \mu^+\mu^-$ events, using the same definitions of probe and passing probes. The systematic uncertainties are estimated by varying the analytical models for the signal and background shapes to fit the invariant mass of the dimuons. Details on the procedure can be found in Ref [146].

The Impact parameter efficiency $\epsilon_{sip|id}$ measurement is performed using $Z \to \mu^+\mu^-$ events. In this measurement, the probe is a muon that satisfies the loose identification criteria, and it is considered a passing probe if it passes the SIP_{3D}, d_{xy}, and d_z cuts of this analysis.

The isolation efficiency $\epsilon_{iso|sip}$ is measured using events from the $Z \to \mu^+\mu^-$ events decay. The isolation of muons is calculated after recovering the FSR photons, described in Section 5.3.3, and after subtracting their contribution to the muon isolation cone. A more detailed description of the method can be found in Ref. [147].

The overall scale factors provided by the data to MC simulation efficiency ratio are shown in Fig. 5.7, estimated as the product of all the data to simulation scale factors for muon tracking, reconstruction, identification, impact parameter, and isolation requirements.

$$\epsilon = \epsilon_{trk} \times \epsilon_{id|trk} \times \epsilon_{sip|id} \times \epsilon_{iso|sip} \tag{5.4}$$

The overall correction is about 1% for most p_T and η values, increasing to about 2% for muons below 10 GeV or with $|\eta| > 2$.

5.3.3 Photons and FSR Recovery

Identifying photons coming from FSR is very important in this analysis for two main reasons: the inclusion of the FSR photons in the calculation of the four leptons invariant mass leads to a sharper

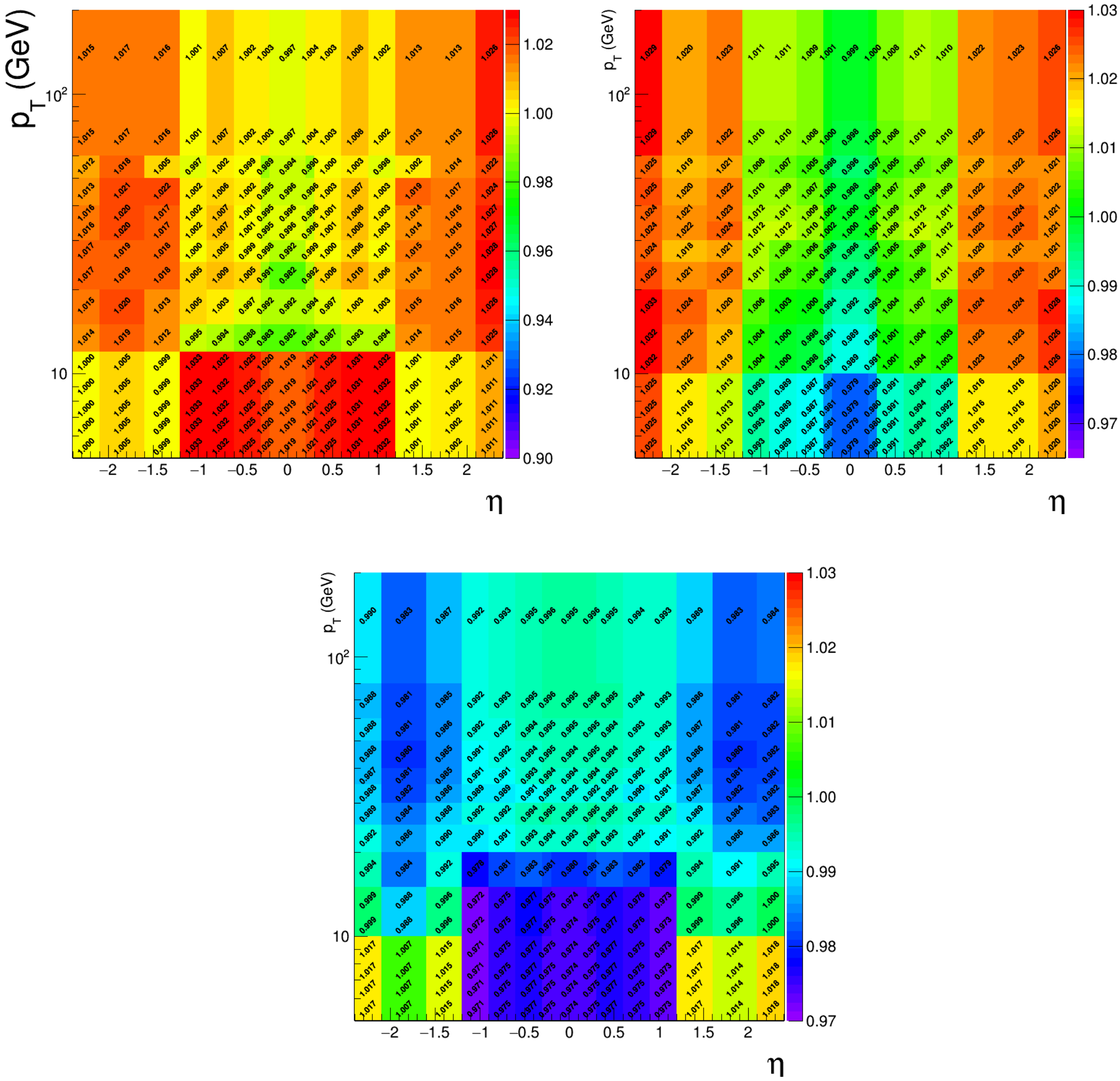


Figure 5.7: Overall data to simulation scale factors for muons, as a function of p_T and η, for 2016 (top-left), 2017 (top-right), and 2018 (bottom). [122]

peak for the Higgs boson signal, in addition, removing FSR photons from the isolation cone of the leptons improves the selection efficiency [148].

The selection of FSR photons is performed for each lepton and does not depend on any Z mass criterion, thus simplifying the subsequent ZZ candidate building and selection. Starting from the collection of photons provided by the PF algorithm (PF photons), the selection of photons and their association to a lepton proceeds as follows:

1. The pre-selection requirements: the PF photons are required to have $p_T^\gamma > 2$ GeV, $|\eta^\gamma| < 2.4$. PF isolation in a cone of size $\Delta R = 0.3$ around the photon is required to be less than 1.8, where the PF isolation for photon is defined as for the muon in equation 5.3. The PF isolation requires a threshold of 0.2 GeV on charged hadrons with a veto cone of 0.0001 and 0.5 GeV on neutral hadrons and photons with a veto cone of 0.01, including the contribution from pileup

vertices (with the same radius and threshold as per-charged isolation).

2. Supercluster veto: All PF photons that match any electron passing both the loose ID and SIP constraints, defined in Section 5.3.1 are removed. The matching is performed by directly associating the two PF candidates.

3. Photons are associated to the closest lepton in the event among all those that pass both the ID and SIP cuts.

4. Photons that do not satisfy $\Delta R(\gamma, l)/E_{T,\gamma}^2 < 0.012$ and $\Delta R(\gamma, l) < 0.5$ are discarded.

5. If more than one photon is associated with the same lepton, we select the one with the lowest $\Delta R(\gamma, l)/E_{T,\gamma}^2$.

6. Each selected FSR photon is removed from the isolation sum for each lepton pass both the loose ID and SIP cuts. This concerns the photons that are in the isolation cone and outside the isolation veto of these leptons.

More details on the optimization of the FSR photon selection can be found in Ref. [146, 147].

5.3.4 Jets

The prominent feature of the searched signal is the presence of two energetic jets originating from light-flavoured quarks, which are typically two valence quarks from each of the colliding protons scattered away from the beamline and inside the detector acceptance by the VBF process. Therefore, these VBF "tagging" jets are roughly expected in the forward region of the detector. The identification, isolation, and energy corrections for jets are described in this section.

5.3.4.1 Jet Identification

The reconstruction of Jets is done by clustering charged and neutral hadrons with the anti-k_T algorithm, explained in Section 2.4.5, with a distance parameter of 0.4. The jet four-momentum is estimated as the vectorial sum of all clustered PF candidates four momenta in the jet and is calculated from simulation to be within 5 to 10% of the true momentum over the entire detector acceptance and p_T spectrum.

Additional proton-proton interactions within the same or nearby bunch crossings can introduce additional tracks and calorimetric energy depositions to the jet. To alleviate this effect, tracks originating from pileup vertices are discarded, and an offset correction is applied to correct the residual contributions.

To account for the jets that may be dominated by reconstruction failures or anomalous contributions from the different subdetector components, we applied additional selection criteria on jets kinematic variables and isolation to be considered in this analysis. Jets are required to have $p_T^{jet} > 30$ GeV, $|\eta^{jet}| < 4.7$ and be separated from all selected lepton candidates and any selected FSR photons by $\Delta R(\ell/\gamma, \text{jet}) > 0.4$. In addition, jets are required to satisfy the tight identification criteria and the tight jet pileup identification working point described in Ref. [149].

5.3.4.2 Jet Energy Corrections

The response of the calorimeter to particles is not linear, and it is not straightforward to convert the measured jet energy to the actual particle or parton energy; therefore, we need a Jet Energy Correction (JEC). The JEC is derived from the simulation to match that of particle level jets on average. In this analysis, standard jet energy corrections are applied to the reconstructed jets as described in Section 2.4.5.

Moreover, momentum balance measurements in Z + jet, photon + jet, dijet, and multijet events are used to account for any residual differences in the jet energy scale in data and simulation. Smearing for the jet energies in the simulation is done to match the data resolution. The jet energy resolution typically amounts to 16% at 30 GeV, 8% at 100 GeV, and 4% at 1 TeV.

5.3.4.3 B-tagging

Jets from the b-quarks hadronization are distinguished using a deep neural network (DeepCSV), which receives the following inputs: jet kinematic variables, displaced tracks from the primary interaction vertex, identified secondary vertices, and related information about the presence of soft leptons in the jet [150]. Three working points are defined: loose, medium, tight, corresponding to 10, 1 and 0.1% probability of misidentifying a light-flavour jet with $p_T > 30$ GeV, respectively.

In this analysis, a jet is considered b-tagged if it passes the medium working point greater than 0.6321, 0.4941, and 0.4184 for 2016, 2017, and 2018, respectively.

Chapter 6

VBF Analysis Strategy

The physics of interest for the present analysis is the VBF Higgs boson production in the $H \rightarrow ZZ \rightarrow 4l$ $(l = e,\ \mu)$ channel. Some SM processes are expected to have a similar final-state topology as the signal and are considered background in the VBF analysis. A typical analysis in particle physics aims to reject these background processes as much as achievable to increase the signal significance. The strategy followed in this analysis is outlined below:

- Most of the collision events from uninteresting processes are removed by the CMS HLT system as discussed in Section 5.1, where dedicated trigger paths are defined in the trigger system to single out the events with the required detector signature for the search presented in this thesis.

- The typical physics objects, which are the result of the VBF interaction, are electrons, muons, and jets. The reconstruction of these objects has been done according to the recommendations from physics object groups as described in Section 5.3.

- Then the selections performed in this analysis are designed to reconstruct a final state with four charged leptons ($4\mu, 4e,$ *or* $2e2\mu$) and at least two jets. The four-lepton system was designed to reconstruct the Higgs boson and has synchronized to the one used to discover the Higgs in the ZZ→4l channel (see Section 6.1).

- A reliable background estimation is performed in order to ensure an accurate assessment of all possible background processes. In this analysis, two types of backgrounds have been considered: "irreducible backgrounds" are estimated from MC simulations and "reducible background" is estimated using a data-driven method, explained in Section 6.2.

- In order to check whether or not the data agrees with the expectation from the simulation, validation plots are made for kinematic distributions of relevant variables, comparing data to the signal and background expectation from simulation. Three control regions were defined to control our observables, defined in Section 6.3.

- The extraction of the VBF signal is done utilizing machine-learning techniques. In order to optimize the analysis and adequately handle the usage of the third jet in the events, two jet-based categories have been defined. For each category, a NN was developed, detailed in Chapter 7.

- A reliable evaluation of the systematic uncertainties that affect the analysis is done, and their impact has been studied, presented in Section 8.1.

- Finally, shape analysis is performed on the VBF ANN discriminants, separately for each jet-based category and their combination. The VBF Higgs boson signal is extracted using the profile likelihood asymptotic approximation, as explained in Section 8.3.

6.1 Event Selection

After selecting all the objects needed to perform the analysis as detailed in Chapter 5, all the stones have to be put together to choose the events of interest. This section will describe the event selection strategy based on a cut-based approach applied to some observables that can separate the VBF Higgs signal from background processes.

6.1.1 Trigger Selection

Most of the collision events from uninteresting processes are removed by the CMS HLT - trigger system, as described in Section 2.3.2, and the dedicated trigger paths are defined in the trigger system to single out the events with the required detector signature for the search presented in this thesis (see Section 5.1). Furthermore, to ensure the full trigger efficiency, the offline p_T threshold for electrons and muons is chosen to be higher than that used in online triggers. For simulated samples, the event is considered when it passes one of the trigger paths. However, for data, double counting of the same event has to be considered. To avoid double-counting, events that have already been selected in another data set are vetoed in a given data set.

6.1.2 Vertex Selection

Vertex reconstruction [63] aims to measure the position of all proton-proton interaction vertices in each event, including the PV and vertices originating from pileup collisions, using the available reconstructed tracks. The PV refers to the position where the hard scattering or proton-proton collisions occur, while the secondary vertices arise from the decay of long-lived particles created from the PV, such as b-jets. The reconstruction of the vertices consists of three steps:

- Selecting tracks consistent with having been generated in the primary interaction region by imposing some requirements on the tracks, such as the track impact parameter relative to the centre of the beam spot, the number of strips, and pixel hits χ^2.

- The clustering of tracks that appear to originate from the same interaction vertex is based on the z-coordinates at their point of closest approach to the centre of the beam spot.

- The fitting of the position of each vertex based on the associated tracks.

In this analysis, the events are required to have at least one good PV fulfilling the following criteria: high number of degree of freedom ($N_{PV} > 4$), collisions restricted along the z-axis ($z_{PV} < 24$ cm) and small radius of the PV ($r_{PV} < 2$ cm).

If more than one primary vertex is reconstructed, the vertex with associated tracks with the highest sum of the square of the transverse momenta is considered as the primary vertex.

6.1.3 ZZ Candidate Selection

The four-lepton candidates are formed from selected leptons, i.e., the tight leptons (defined in Sections 5.3.1 and 5.3.2) that satisfy the SIP_{3D} vertex constraint and the isolation cuts. FSR photons are subtracted as described in Section 5.3.3. In addition, lepton cross-cleaning is performed by discarding electrons within $\Delta R < 0.05$ of selected muons.

The construction and selection of four-lepton candidates proceed, as the one used to discover the Higgs in the ZZ→4l channel, according to the following sequence:

1. **Z candidates:** are defined as all possible pairs with opposite charges and the same flavour of selected leptons (e^+e^-, $\mu^+\mu^-$) that satisfy $12 < m_{ll(\gamma)} < 120$ GeV/c^2, including the FSR photon if it exists. The di-lepton invariant mass distributions for each year of Run 2 are shown in Fig. 6.1. Data is represented by black dots, while the different backgrounds that affect the analysis in this phase space are shown in coloured histograms. The dominant background at this step of the analysis is the DY process. The plots show a good agreement between data and MC.

2. **ZZ candidates:** are built from a pair of Z candidates that do not share the same lepton (non-overlapping). The Z candidate with invariant mass $m_{ll(\gamma)}$ closest to the nominal Z boson mass is indicated as Z_1 and the second one is the Z_2. This step eliminates most of the contribution coming from Z+jets, WZ and $t\bar{t}$ backgrounds. The reconstructed ZZ candidates are required to satisfy the following list requirements:

   (a) **Ghost removal**: any two leptons must satisfy $\Delta R(\eta, \phi) > 0.02$ to avoid reconstructing a single object as two or more objects.

   (b) **Lepton p$_T$**: at least two out of the four selected leptons must have $p_T^1 > 10$ and $p_T^2 > 20$ GeV.

   (c) **QCD suppression**: all opposite-sign lepton pairs that can be formed from the four leptons (regardless of the lepton flavour) must satisfy $m_{ll} > 4$ GeV. Here, the selected FSR photons are not included in the mass calculation since QCD-induced low mass dilepton resonances, such as J/Ψ, may have nearby photons from the π^0 decay.

   (d) **Z_1 invariant mass**: $m_{Z_1} > 40$ GeV.

   (e) **Smart cut**: defining Z_a and Z_b as the mass-sorted alternative pairing Z candidate (Z_a being the closest one to the nominal Z mass), require NOT($|m_{Z_a} - m_Z| < |m_{Z_1} - m_Z|$ AND $m_{Z_b} < 12$). Selected FSR photons are included in the m_Z's calculation. This cut discards 4μ and 4e candidates which have the alternative pairing similar to an on-shell Z + a low-mass pair l^+l^-.

   (f) **Four-lepton invariant mass**: $m_{4l} > 70$ GeV.

3. **Choice of the best ZZ candidate**: if more than one ZZ candidate survives the previous selections, we choose the pair with m_{Z1} closest to the nominal Z mass and m_{Z2} with the largest scalar sum of the leptons p_T.

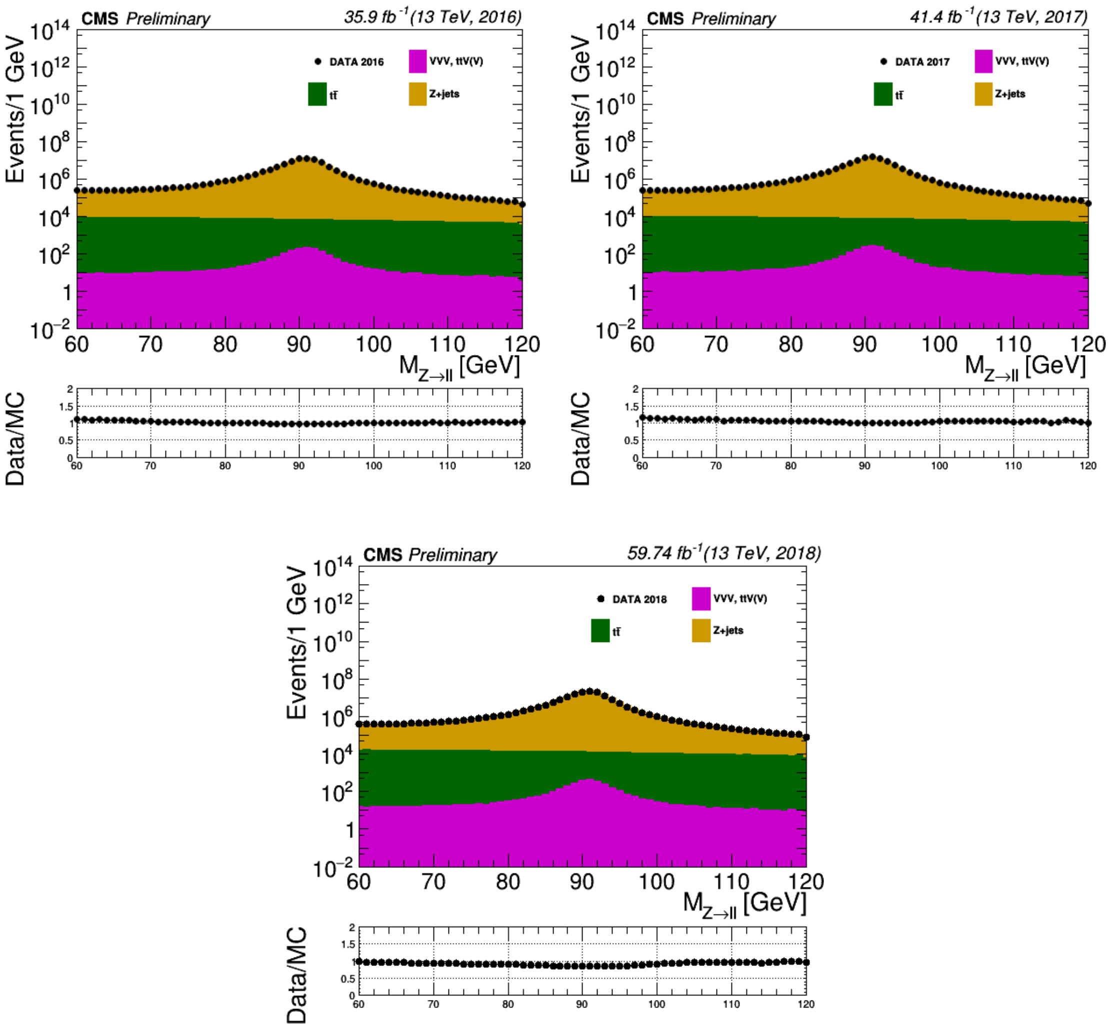


Figure 6.1: Di-lepton invariant mass distributions for 2016 (top-left), 2017 (top-right), and 2018 (bottom). Data represented by black dots and simulation represented by coloured histograms.

The reconstructed four-lepton invariant mass distribution is shown in Fig. 6.2 for the full Run 2 dataset, compared to the expected backgrounds, for the full mass range and zooming on the low-mass range. In Fig. 6.3, the same distributions are shown for each year of Run 2 for the same two mass ranges.

Events containing at least one ZZ candidate satisfying all the previous selections are used to tag the Higgs decaying into a four-lepton final state. In order to enhance the presence of VBF events over the other SM Higgs production modes, additional cuts are applied, as described in the next section.

6.1.4 VBF Signal Region

In order to optimize the analysis for the VBF production mode, further requirements are imposed on the events passing the selection criteria described in the previous section. These additional selections are:

- **Number of jets**: exploiting the existence of two energetic jets in the final state of our signal

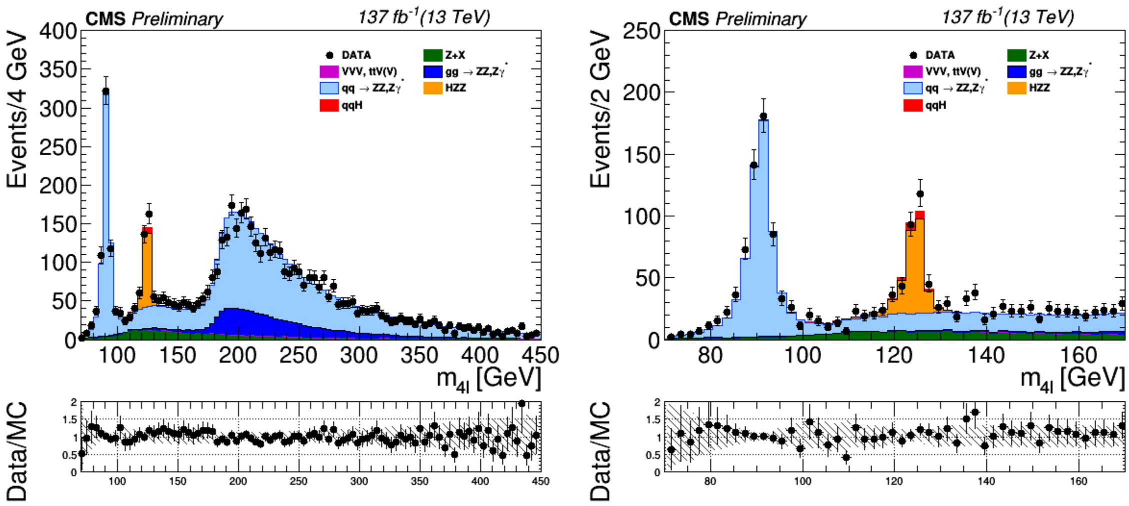


Figure 6.2: The four-lepton invariant mass distributions after the full four leptons selection for the combined Run 2 for the full mass range (right) and low mass range (left). The distributions include the $Z+X$ background estimated by the data-driven method explained in Sec. 6.2.2.

that can be used to tag the events; hence we require the event to have

- EITHER, two or three jets from which at most one b-tagged jet;
- OR, more than three jets with no b-tagged jet;

passed the jet identification criteria described in Section 5.3.4.

- **Four-lepton invariant mass**: events must have $118 \leq m_{4l} \leq 130$ GeV. This cut is powerful to discriminate the VBF signal from the SM backgrounds, such as the non-resonant ZZ background.

- **Number of leptons**: requires exactly four leptons passed the tight selection criteria described in Sections 5.3.1, 5.3.2. This cut reduces the contribution of processes with more than four real leptons in the final state, such as $t\bar{t}$+VV(V), VVV, ZH, WH, and $t\bar{t}$tH.

The four-lepton invariant mass distributions for the events passing the VBF signal-region selections are shown in Fig. 6.4. The Z+X background is estimated by the data-driven method explained in Section 6.2.2.

6.2 Background Estimation

A reliable background estimation is crucial in any physics analysis to ensure an accurate assessment of all possible background processes. In this analysis, two types of backgrounds have been considered: "irreducible background", which represents all the processes that yield the same final state as the signal, and "reducible background", which represents all the processes that do not have the same final state as the signal process but can mimic the final state of the signal due to the misreconstruction of physics objects in the detector.

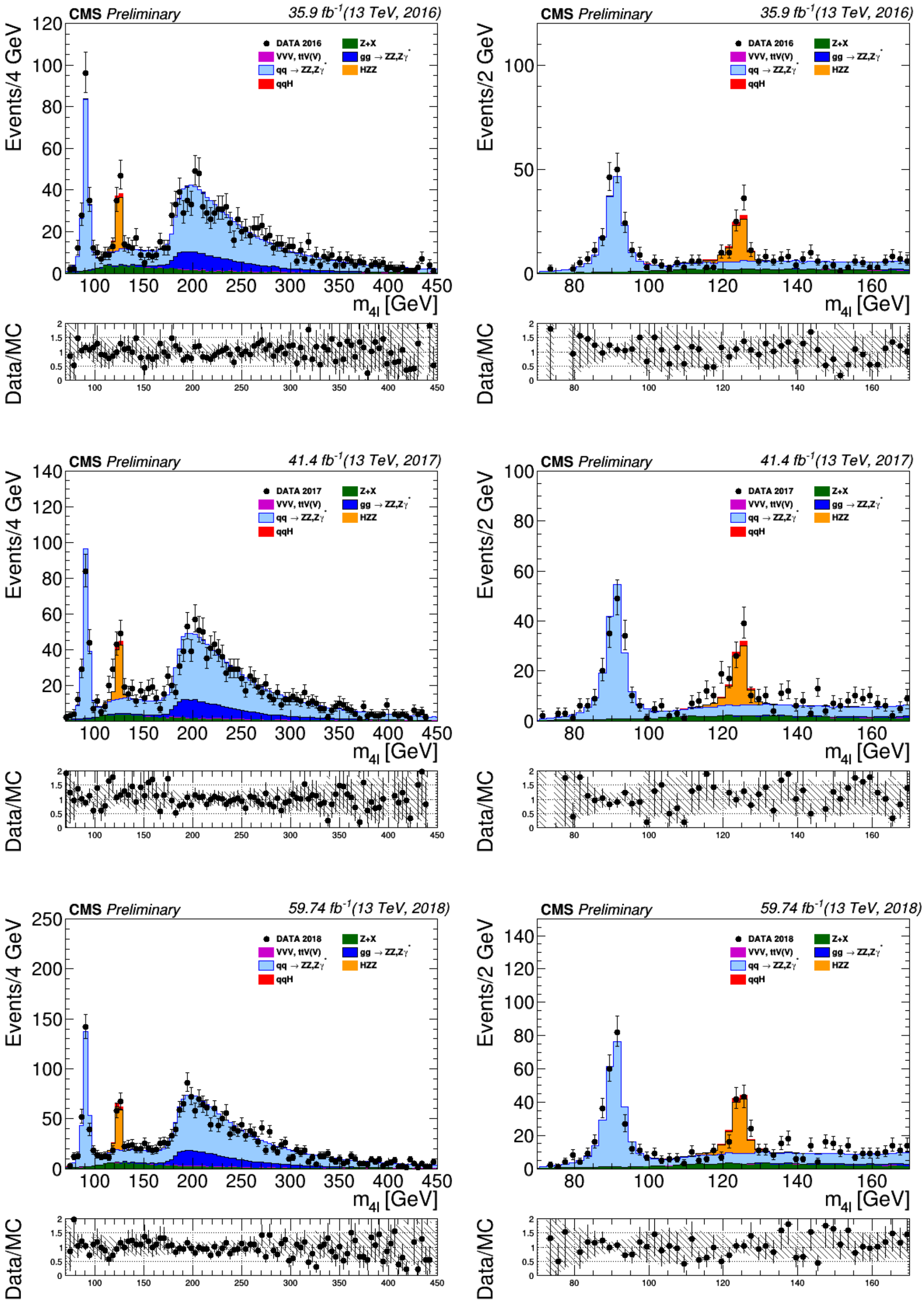


Figure 6.3: The four-lepton invariant mass distributions after the full four leptons selection for 2016 (top), 2017 (middle), and 2018 (bottom) datasets for the full mass range (right) and low mass range (left). The distributions include the $Z + X$ background estimated by the data-driven method explained in Sec. 6.2.2.

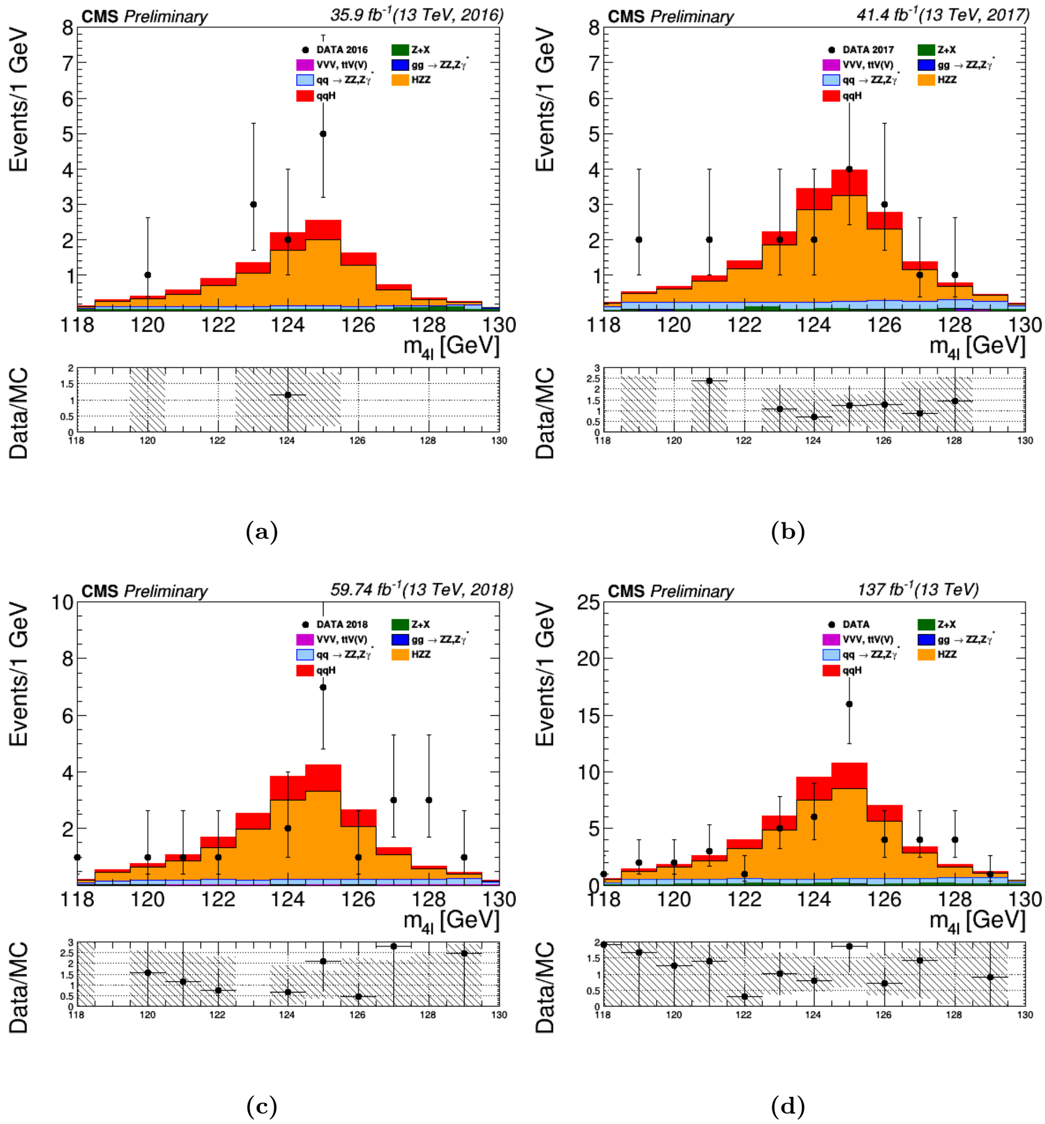


Figure 6.4: Final four-lepton mass distributions for the events passing the VBF-SR selections (a) 2016, (b) 2017 and (c) 2018 and, (d) the combined Run 2. included the $Z + X$ background estimated by the data-driven method explained in Sec. 6.2.2.

6.2.1 Irreducible Backgrounds

The irreducible backgrounds for this analysis are composed by the three SM Higgs production modes, gluon fusion (ggH) and associated vector (WH, ZH and $t\bar{t}H$), and the SM backgrounds $q\bar{q} \to ZZ$, gg $\to ZZ$. These backgrounds are estimated from MC simulations, as revealed in this section.

- **SM Higgs Production Modes**
  The SM Higgs production modes (ggH, WH, ZH, and $t\bar{t}H$), decaying to ZZ bosons, represent the dominant background for this analysis. They are estimated from simulation by normalizing the yields according to the cross-sections times the branching ratio fractions (listed in Table. 5.2).

- **Modeling of $\mathbf{q\bar{q} \to ZZ}$**

The $q\bar{q} \to ZZ$ background events are generated at NLO accuracy. The fully differential cross-section is computed at NNLO, which is not yet available in a partonic level event generator. Consequently, NNLO/NLO k-factors are applied to the $q\bar{q} \to ZZ$ NLO POWHEG sample as function of m_{ZZ} [122]. Additional NLO electroweak corrections, which depend on the initial state quark flavour and kinematics, are also applied in the mass range of $m_{ZZ} > 2m_Z$, where the corrections have been computed [122].

- **Modeling of gg → ZZ**
  The $gg \to ZZ$ background is generated at LO with the generator MCFM 7.0 [133]. This background does not have exact calculations beyond LO. However, it has shown that the soft collinear approximation can describe such background cross-section and the interference term at NNLO [151]. In addition, the k-factors are very similar at NLO for $gg \to H$ and $gg \to ZZ$, and at NNLO $gg \to H$ and the interference terms [152]. Because of that, the same k-factor is used for both samples.

6.2.2 Reducible Background

The system commonly referred to as Z+X is a reducible background in the VBF $H \to ZZ \to 4l$ analysis. It originates from processes containing one or more non-prompt leptons in the four-lepton final state. Non-prompt leptons are mainly non-isolated leptons that can originate from the decay of heavy-flavour mesons, misreconstructed jets (usually originating from light-flavour quarks) and electrons from γ conversions. Such non-prompt leptons are usually referred to as “fake” leptons. The main processes that lead to these backgrounds are Z+jets, $t\bar{t}$, WZ, ZZ, $Z\gamma$ production. The estimation of such a background in the VBF $H \to ZZ \to 4l$ analysis is done by measuring the f_e (f_μ) probability that fake electrons (fake muons) passing the *loose* selection criteria to also pass the final selection criteria (the *tight* selection). These probabilities are called fake ratios or fake rates and are applied to two defined control regions (CRs) to extract the expected Z+X yield in the signal region. The following sections describe the steps required to estimate the Z+X yield and shape in this analysis.

6.2.2.1 Measuring the Fake Rates

To measure the leptons fake rates, we build a $Z(\ell\ell)$ + e + at least 2 isolated jets, and $Z(\ell\ell) + \mu$ + at least 2 isolated jets control regions that are expected to be completely dominated by final states which include a Z boson, a fake lepton and at least two jets isolated with respect to the leptons.

The events that populate the defined control regions should satisfy the following requirements:

- The leptons forming the Z must be an opposite sign, same flavour and have $p_T > 20(10)$ GeV. The FSR photon is included in Z reconstruction if it exists.

- The third lepton (e/μ) has to pass the loose selection (kinematic and impact parameter cuts); this lepton is used as the probe to compute the f_e and f_μ probabilities.

- The invariant mass given by the fake lepton and the opposite-sign tight lepton composing the Z is required to be $m_{ll} > 4$ GeV to reduce QCD contamination.

- In order to suppress contamination from γ conversions to electrons, the invariant mass of the two leptons composing the Z must satisfy $|m_{Z_{ll}} - m_Z| < 7$ GeV.
- ΔR > 0.02 between any of the three leptons.
- The missing transverse energy in the event is required to be $E_T^{miss} < 25$ GeV to suppress the contribution from the WZ process, which has three real leptons in the final state.

The fake rate is parameterized in terms of the probe lepton p_T and η. Figs. 6.5 - 6.7 show the fake rate estimation separately for electrons and muons in terms of p_T and η. The points with solid error bars are the fake rates computed directly from the procedure described above. As the WZ background potentially contributes with three real leptons and such background is already included in the analysis via MC, it needs to be subtracted from the estimation. The resulting FR after this subtraction is shown by the points with dotted error bars.

With the requirement of two isolated jets, we force the probe lepton to be far away from jets, eliminating most of the events with a fake object coming from the surroundings of the jets. Consequently, lowering the fake rate probability, especially at low p_T, where most bjets are miss-reconstructed as leptons.
Now, as one can see, there is a significant dependence between the FR's and the (p_T, η) of the loose leptons. For this reason the final parameterization of the FR's is made as a function of these two variables and it is shown in the Figs. 6.8 - 6.10. The muon case is worse than electron due to low statistics in data. The values of the fake rates are used to weigh the events populating the control regions defined in the next section.

6.2.2.2 Building Control Regions

In order to apply the previously described fake rates, two control regions are defined, requiring events with two leptons passing the selection of the first Z (tight leptons) and a pair of loose leptons of the same flavour, opposite charge and passing the loose selection cuts. In addition, we require the existence of two jets that satisfy the identification criteria for jets and are isolated from the other objects in the event, as described in Section 5.3.4. These events must also satisfy all the kinematic cuts applied for the selection of the Higgs phase space (see Section 6.1).

The first control region is then built by requiring that the two loose leptons do not pass the final identification and isolation criteria. This control sample is occupied with two Prompt leptons, two Fail leptons, and two jets, referred to as “2P2F” from now on. It is expected to be populated by events that have only two prompt leptons: mostly Z+jets, $t\bar{t}$ and $Z\gamma$.

The second control region is obtained by requiring that just one of the four leptons does not pass the final identification and isolation criteria. The remaining three leptons in the event should pass those selections. This control sample is occupied with Three Prompt leptons, one Fail lepton, and two jets, referred to as “3P1F” from now on. The events appearing in this control region are the same in the 2P2F, with different relative proportions, plus WZ events, which are expected to have three prompt leptons.

These two control regions, enriched by fake leptons, are orthogonal to the Higgs signal region and are used to estimate the yields and shape of the Z+X background in the signal region.

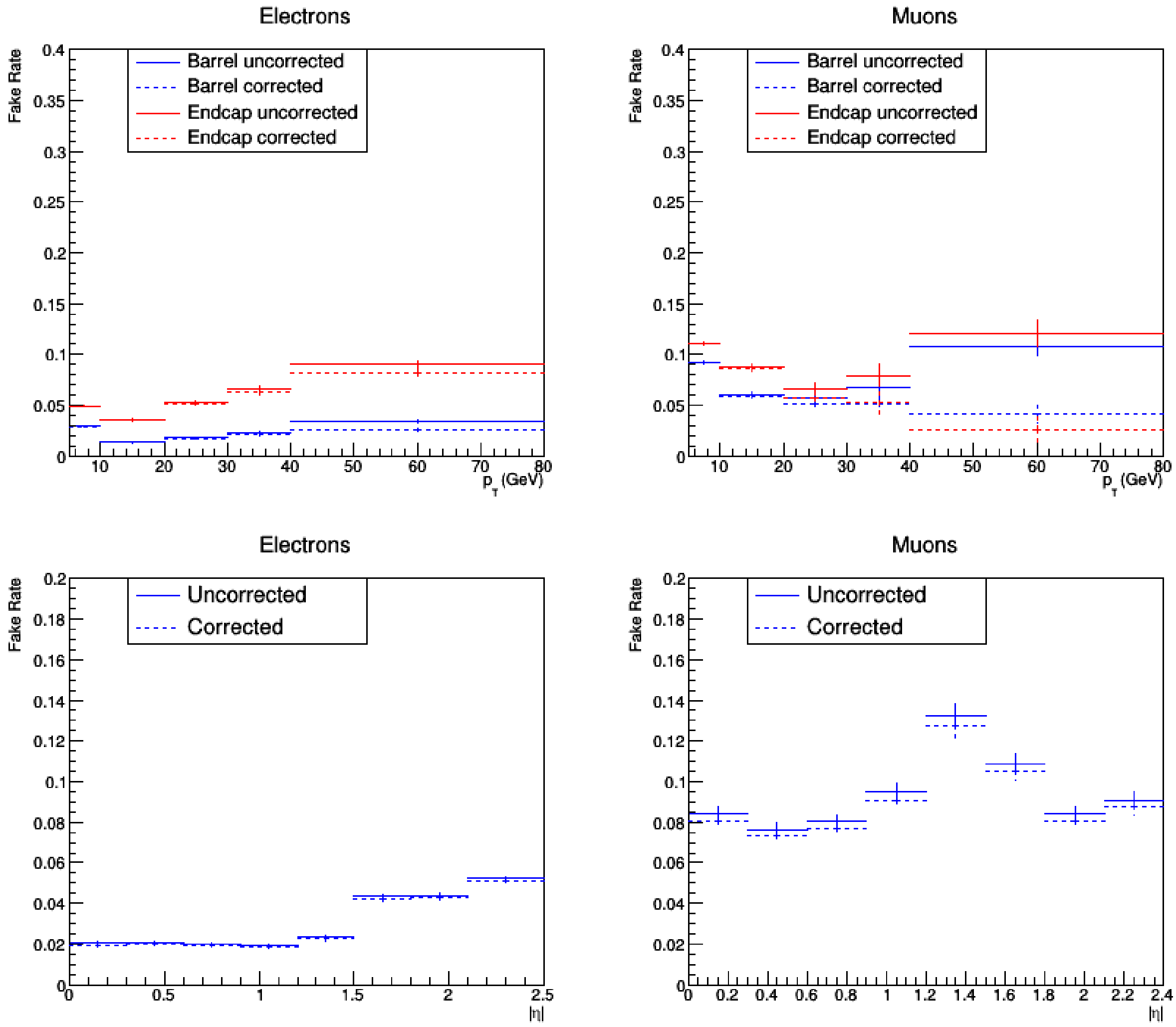


Figure 6.5: Fake rates for 2016 dataset are shown in terms of the lepton probe p_T (top) and η (bottom). The p_T distributions (barrel and endcap) corresponds to the regions defined by $|\eta| < 1.479$ (1.2) for electrons (muons). The fake rates are shown before (solid lines) and after (dotted line) removal of WZ contribution from MC.

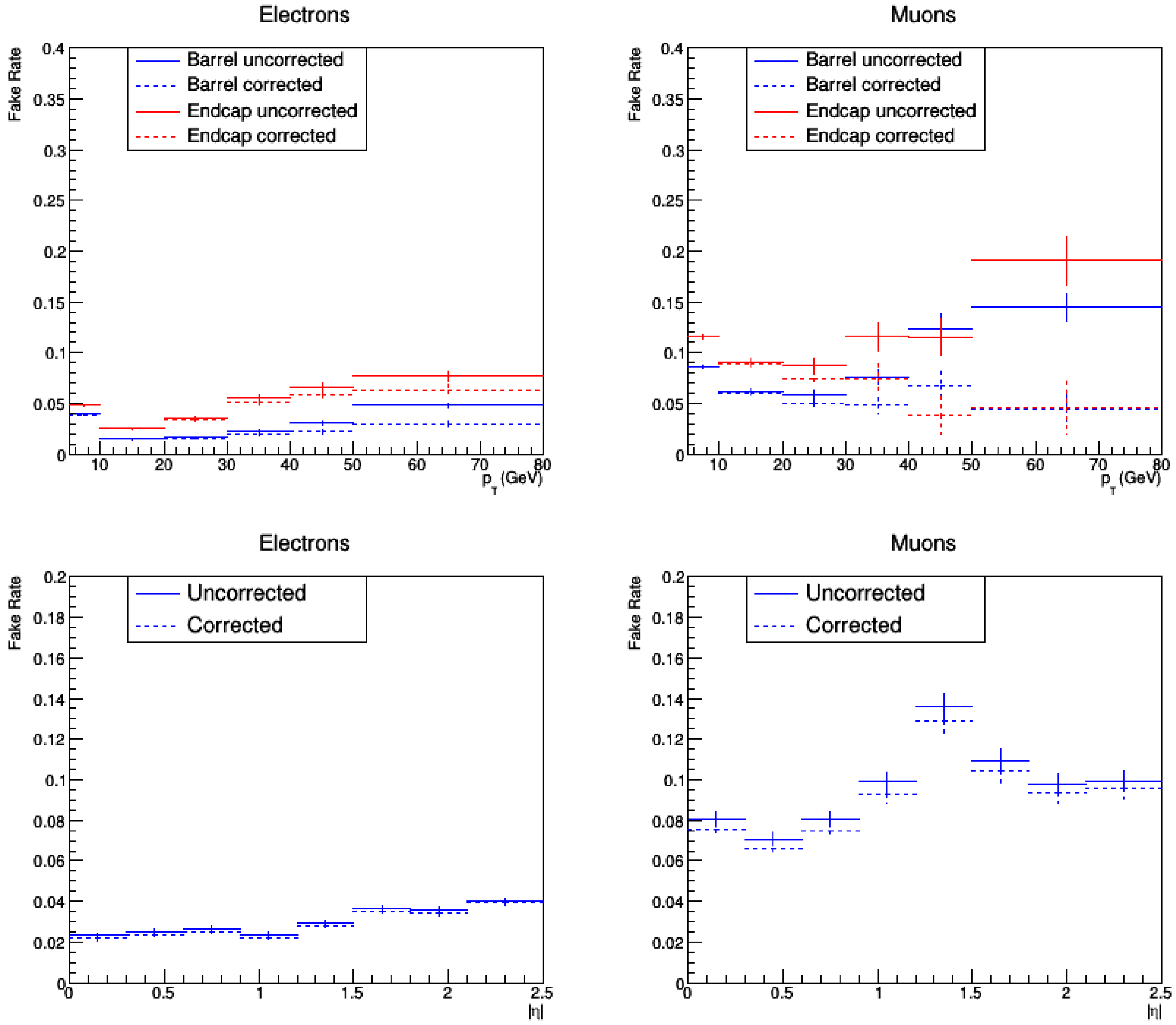


Figure 6.6: Fake rates for 2017 dataset are shown in terms of the lepton probe p_T (top) and η (bottom). The p_T distributions (barrel and endcap) corresponds to the regions defined by $|\eta| < 1.479$ (1.2) for electrons (muons). The fake rates are shown before (solid lines) and after (dotted line) removal of WZ contribution from MC.

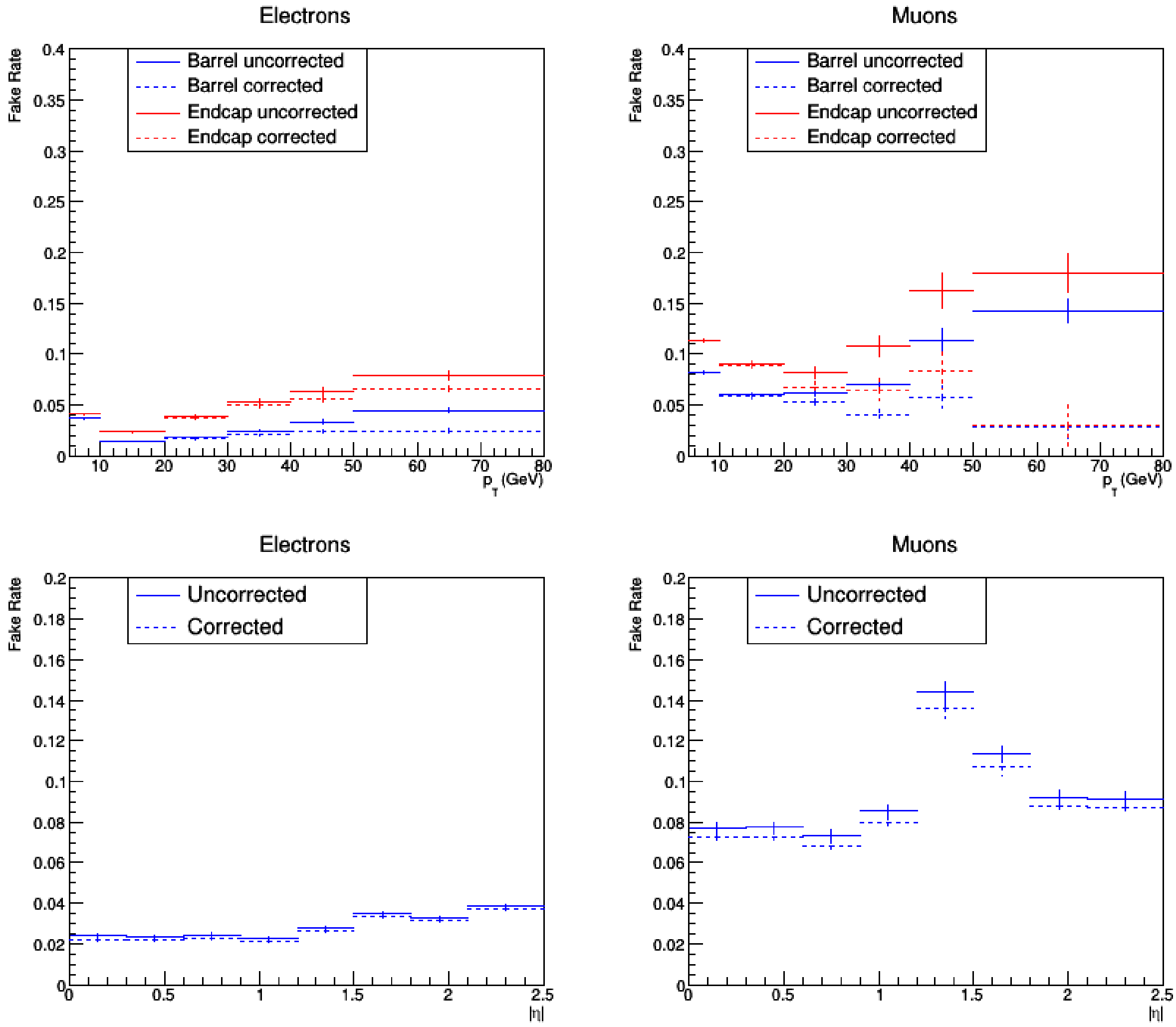


Figure 6.7: Fake rates for 2018 dataset are shown in terms of the lepton probe p_T (top) and η (bottom). The p_T distributions (barrel and endcap) corresponds to the regions defined by $|\eta| < 1.479$ (1.2) for electrons (muons). The fake rates are shown before (solid lines) and after (dotted line) removal of WZ contribution from MC.

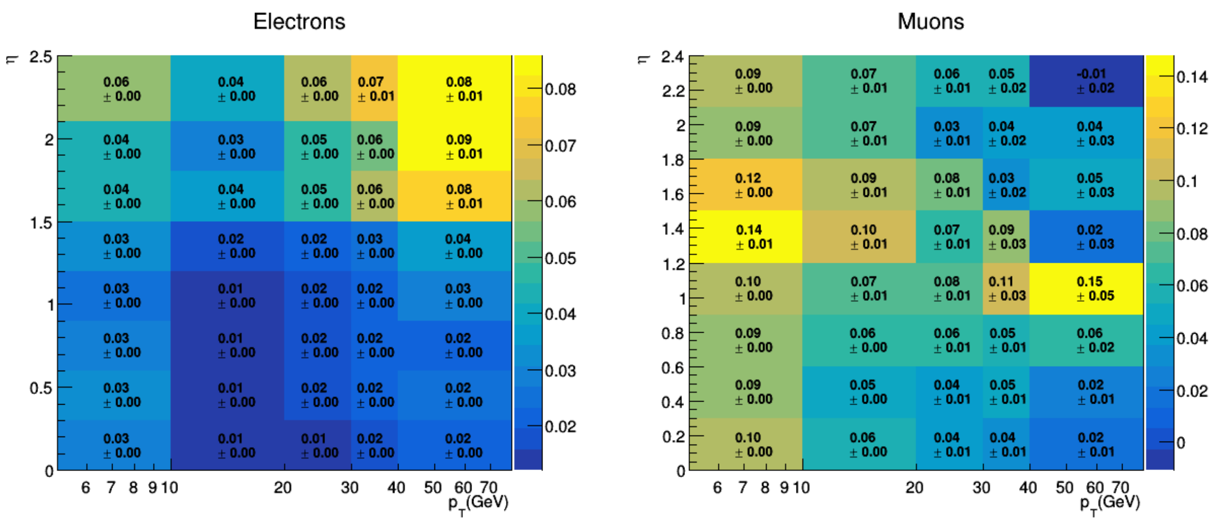


Figure 6.8: Fake rates for 2016 dataset as a function of the probe lepton p_T and η. The background WZ from MC has been subtracted. This is the mapping for fake rates applied into the control regions.

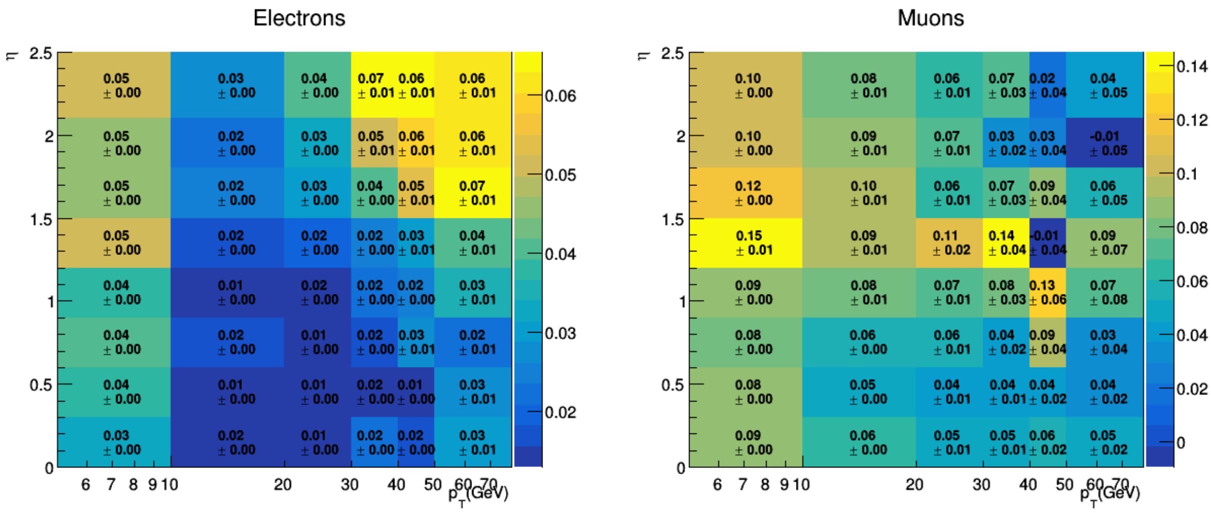


Figure 6.9: Fake rates for 2017 dataset as a function of the probe lepton p_T and η. The background WZ from MC has been subtracted. This is the mapping for fake rates applied into the control regions.

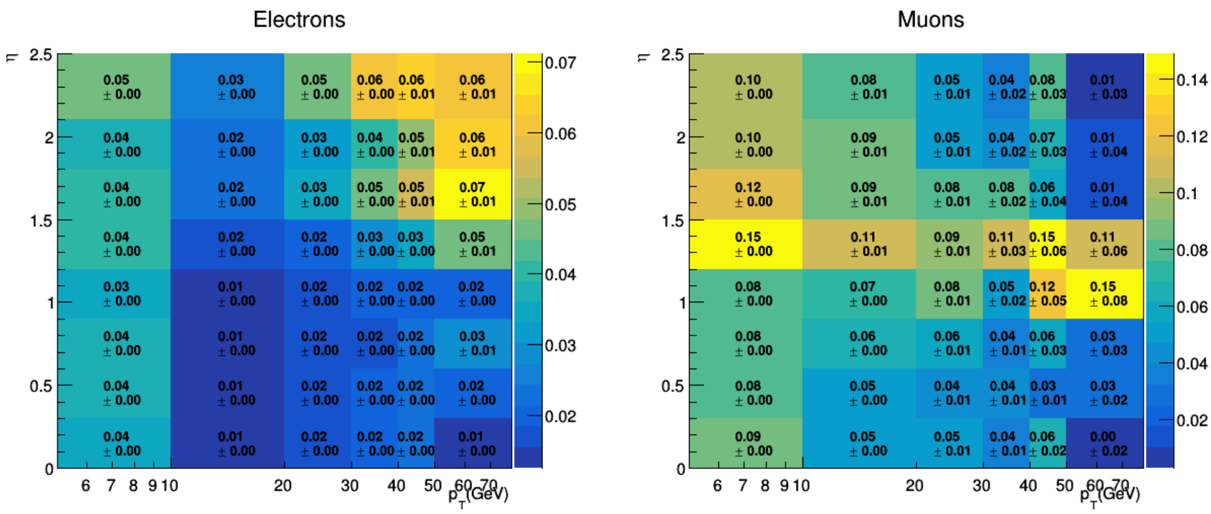


Figure 6.10: Fake rates for 2018 dataset as a function of the probe lepton p_T and η. The background WZ from MC has been subtracted. This is the mapping for fake rates applied into the control regions.

Channel	4μ	4e	$2e2\mu$	$2\mu 2e$	4l
2016	0.29±0.05	0.10±0.05	0.19±0.04	0.02±0.02	0.61±0.09
2017	0.20±0.05	0.04±0.03	0.18±0.05	0.03±0.02	0.44±0.08
2018	0.23±0.05	0.06±0.04	0.21±0.05	0.09±0.06	0.58±0.10

Table 6.1: Yields estimated for Z+X in the signal region from the measurements on data using the fake rate method. Estimates reported with statistical uncertainties.

6.2.2.3 Final Z+X estimation

The expected number of reducible background events in the 3P1F control region can be computed by weighting each event observed in the 2P2F control region with the factor $(\frac{f_i}{1-f_i} + \frac{f_j}{1-f_j})$, in which the f's correspond to fake rates of each loose lepton:

$$N_{\mathrm{3P1F}}^{bkg} = \sum(\frac{f_i}{1-f_i} + \frac{f_j}{1-f_j})N_{\mathrm{2P2F}} \tag{6.1}$$

If the fake rates were measured in a sample having the same background composition as the $2P+2F$ sample, the difference between the observed number of events in the $3P+1F$ sample and the expected background predicted from the $2P+2F$ sample would only account for the (small) contribution of WZ and $Z\gamma_{conv}$. Significant differences arise because the fake rates used in equation 6.1 do not properly account for the background composition of the $2P+2F$ control sample.

In order to correct those discrepancies and correctly estimate the Z+X background, one needs two components:

- a component directly from 2P2F, computed by weighting each observed event in the 2P2F control region using the factor $(\frac{f_i}{1-f_i} \frac{f_j}{1-f_j})$, in which the f's are the fake rates of loose leptons as explained before;
- a component from 3P1F, obtained by the difference between the observed events in 3P1F (N_{3P1F}), the expected contribution from 2P2F and from ZZ in the signal region ($N_{3P1F}^{ZZ} + N_{3P1F}^{bkg}$). The ZZ contribution is taken from the MC simulated events selected in the 3P1F control region. The N_{3P1F}^{bkg} is the 3P1F extrapolation from 2P2F as discussed above.

The final expression for the estimation of Z+X can be then symbolically represented as,

$$N_{SR}^{bkg} = \sum \frac{f_a^i}{(1-f_a^i)}(N_{3P1F} - N_{3P1F}^{bkg} - N_{3P1F}^{ZZ}) + \sum \frac{f_b^i}{(1-f_b^i)} \frac{f_c^j}{(1-f_c^j)} N_{2P2F} \tag{6.2}$$

which is more conveniently expressed as,

$$N_{SR}^{bkg} = (1 - \frac{N_{3P1F}^{ZZ}}{N_{3P1F}}) \sum_i^{N_{3P1F}} \frac{f_a^i}{(1-f_a^i)} - \sum_j^{N_{2P2F}} \frac{f_b^j}{(1-f_b^j)} \frac{f_c^j}{(1-f_c^j)}. \tag{6.3}$$

Equation 6.3 represents what has been done to estimate the Z+X background contribution in this analysis. The Z+X event yield in the signal region is obtained after applying the additional cuts of the selection; yields obtained are shown in Table. 6.1.

6.2.2.4 Uncertainties on the estimation of Z+X

The statistical uncertainty on the Z+X estimation comes from the limited number of events selected in the control regions where one measures and applies the fake ratio methods. The systematic uncertainty arises because the composition of reducible backgrounds in the control regions where one measures and applies the fake ratios is usually not the same, representing the main source of systematic uncertainty of the fake ratio method. This uncertainty is estimated in the following steps: First, measuring the fake rates for each background process (Z+jets, $t\bar{t}$, WZ, ZZ, $Z\gamma$) using MC and applying the selections explained in the Section 6.2.2.1 to obtain samples of $Z + l$. Then, one defines the fake rates for simulation by computing the weighted average of those individual fake rates. Second, one reweights the individual fake rates using the expected yields of the reducible backgrounds selected in the 2P2F control region. Then, one applies the average and the reweighted fake rates in order to get the final yields of Z+X and finally, the difference between them is used as an estimation of the uncertainty on the measurement of fake rates. This uncertainty amounts to about 30 - 41 % [122], depending on the final state and each year of data taken.

6.3 Control regions

To check whether or not the data agrees with the expectation from the simulation, validation plots are made for kinematic distributions of relevant variables, comparing data to the signal and background expectation from simulation. Several observables are studied at different selection steps of the analysis; mainly, three control regions were defined to control our observables:

- Z + at least two jets, this control region is built by requiring the presence of two leptons forming a Z candidate and at least two jets, whose selection is added on top of the Z selection. This control region is chosen to study the agreement between data and simulation for jet observables used in the NN training (described in Chapter 7) in a region with enough statistics up to the 3^{rd} jet.
- $m_{4l} < 118$ or $m_{4l} > 130$ GeV, The four-lepton sidebands region is built with excluding the mass window of the signal. The most important source of background in this region is the ZZ non-resonant production; hence this control region helpful to study the lepton variables.
- $m_{4l} < 118$ or $m_{4l} > 130$ GeV + at least two jets, this sideband region is helpful to double-check the jets observables in the same phase space of the analysis.

Figs. 6.11 - 6.13 show the transverse momentum of the leading jet (top), the next to leading jet (middle), and the 3^{rd} jet (bottom) for 2016, 2017, and 2018, respectively; the distributions in the control region Z + at least two jets shown on the left and the sideband region + at least two jets on the right.

The quark-gluon likelihood of the leading jet is shown in Fig 6.14 in the control region Z + at least two jets (left) and the sideband region + at least two jets (right) for 2016 (top), 2017 (middle), and 2018 (bottom). The four leptons transverse momentum distribution is reported in the sideband region in Fig. 6.15 for 2016 (top left), 2017 (top right) and 2018 (bottom).

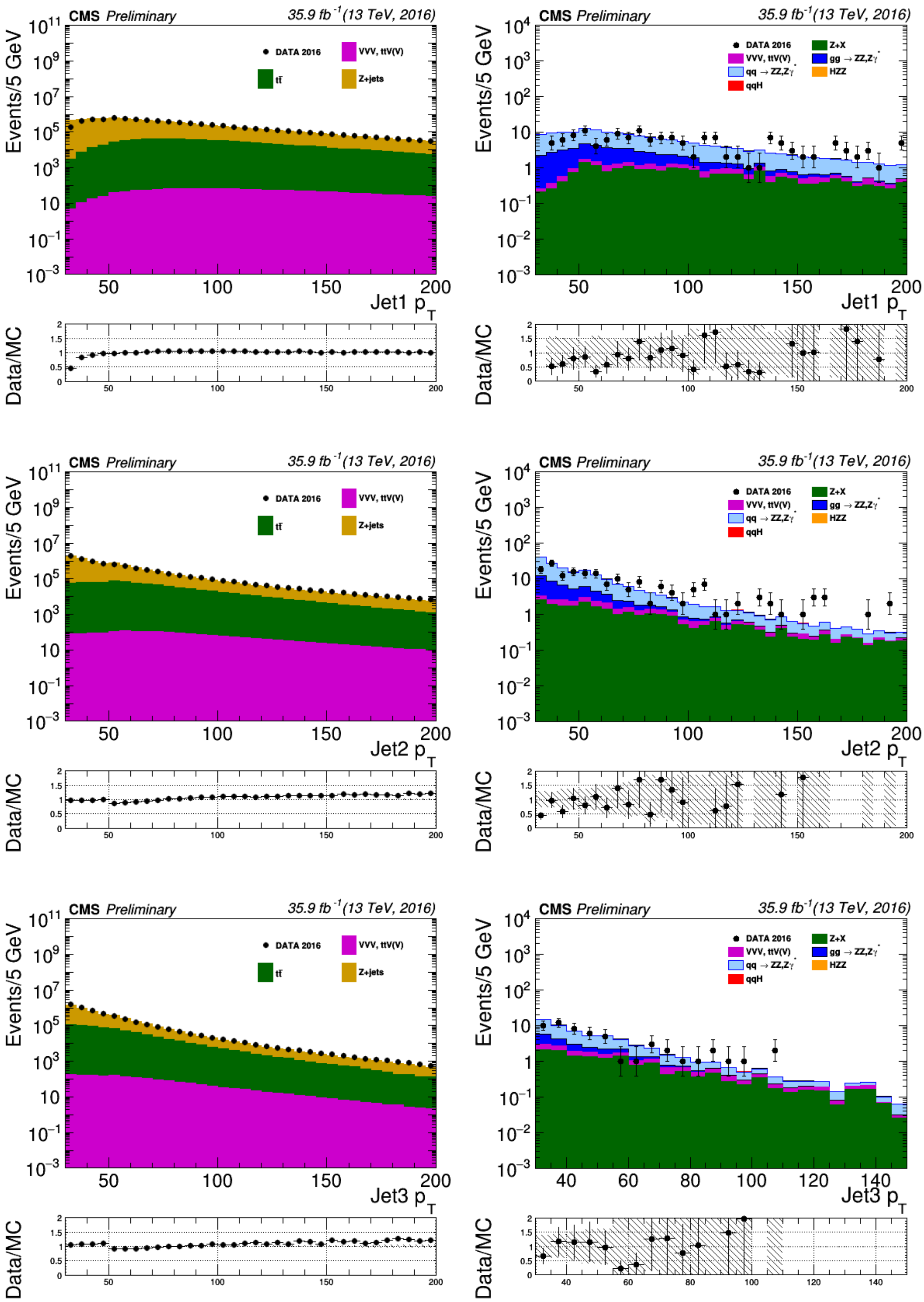


Figure 6.11: Transverse momentum of the jets for the 2016 dataset in the control region Z + at least two jets (left) and the sideband region + at least two jets (right), for the leading jet (top), the next to leading jet (middle), the 3^{rd} jet (bottom).

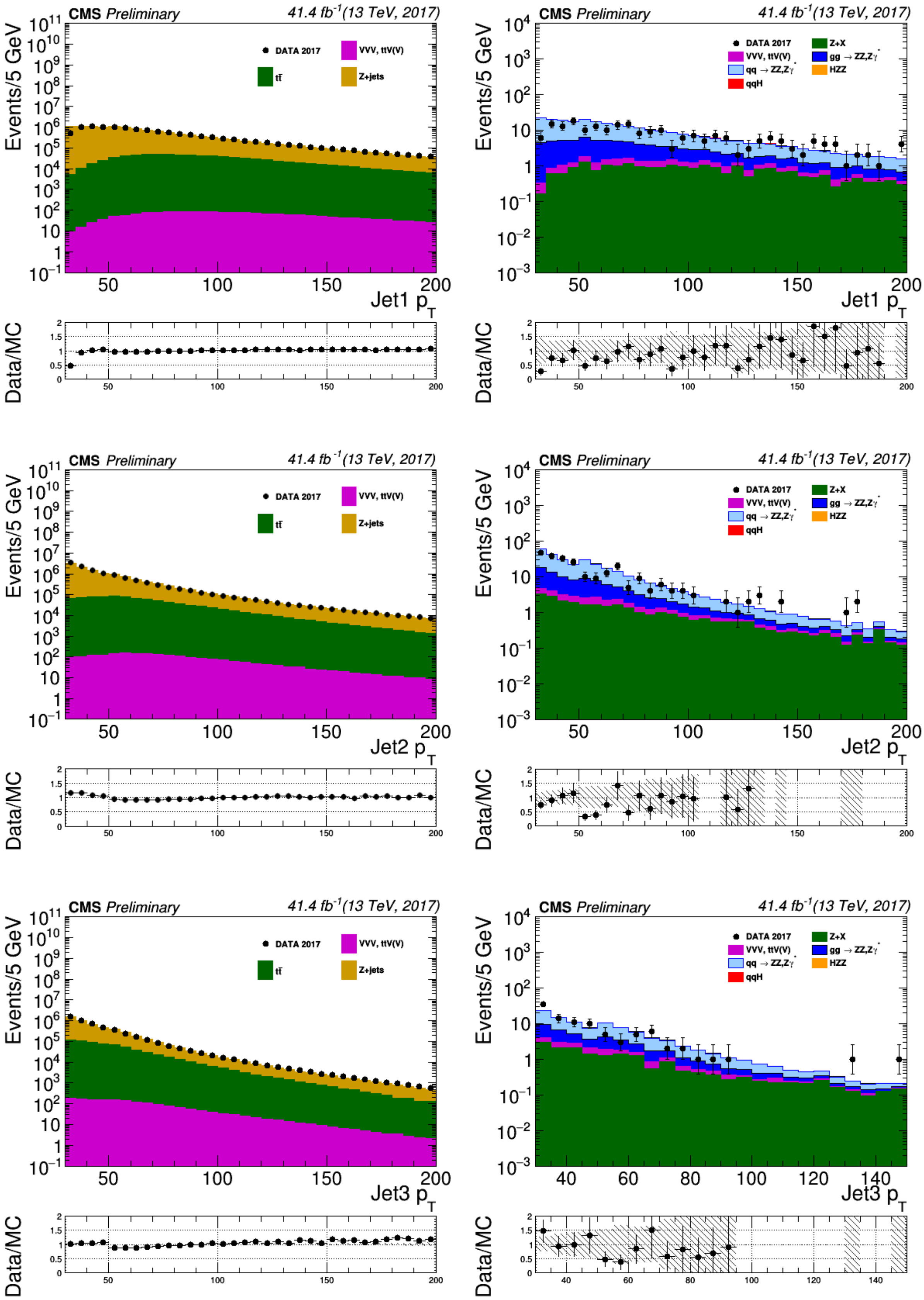


Figure 6.12: Transverse momentum of the jets for the 2017 dataset in the control region Z + at least two jets (left) and the sideband region + at least two jets (right), for the leading jet (top), the next to leading jet (middle), the 3^{rd} jet (bottom).

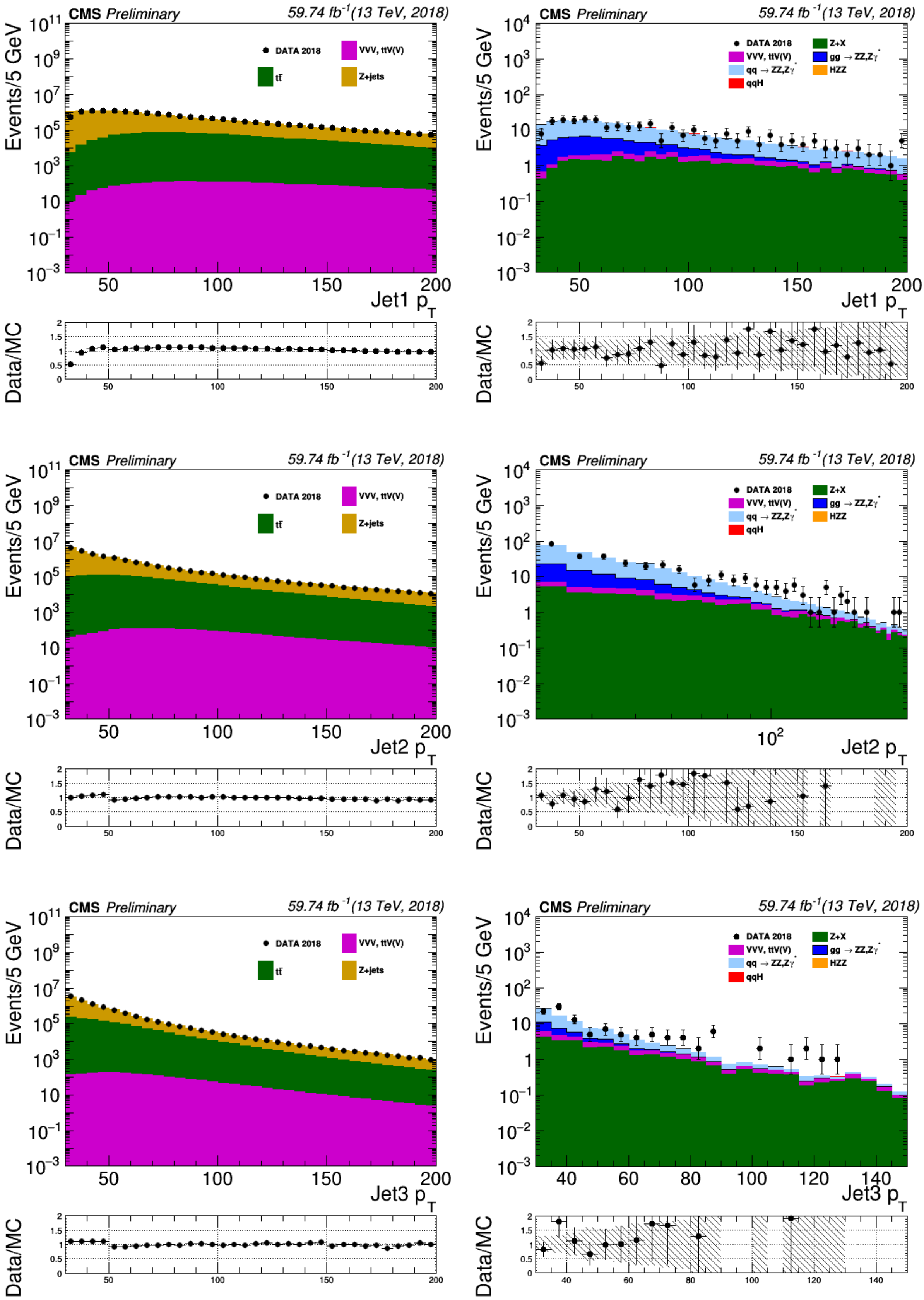


Figure 6.13: Transverse momentum of the jets for the 2018 dataset in the control region Z + at least two jets (left) and the sideband region + at least two jets (right), for the leading jet (top), the next to leading jet (middle), the 3^{rd} jet (bottom).

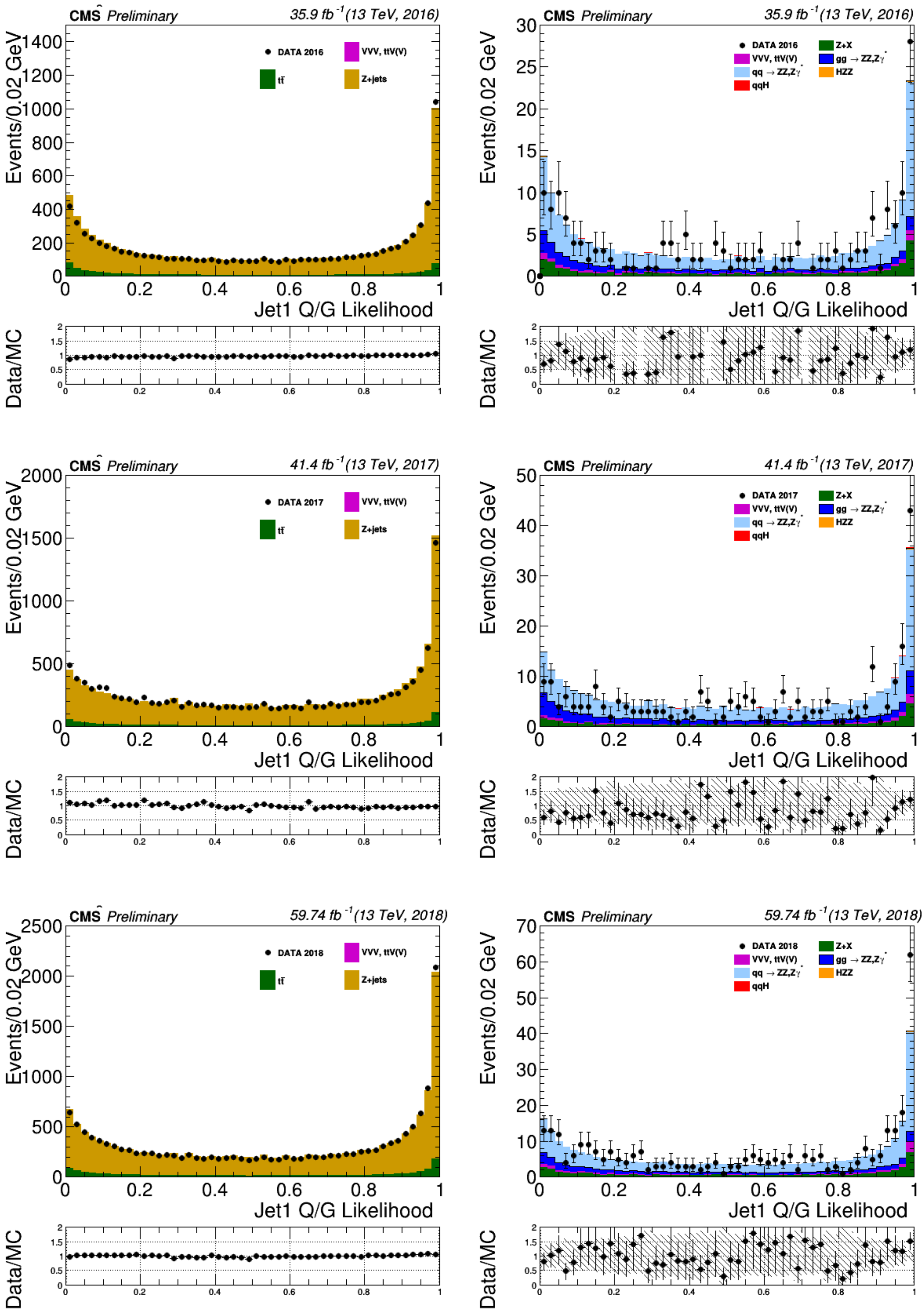


Figure 6.14: The quark-gluon likelihood of the leading jet in the control region Z + at least two jets (left) and the sideband region (right) for 2016 (top), 2017 (middle), and 2018 (bottom).

As can be seen from the plots, there is a good agreement between data and MC for the main observables used in this analysis.

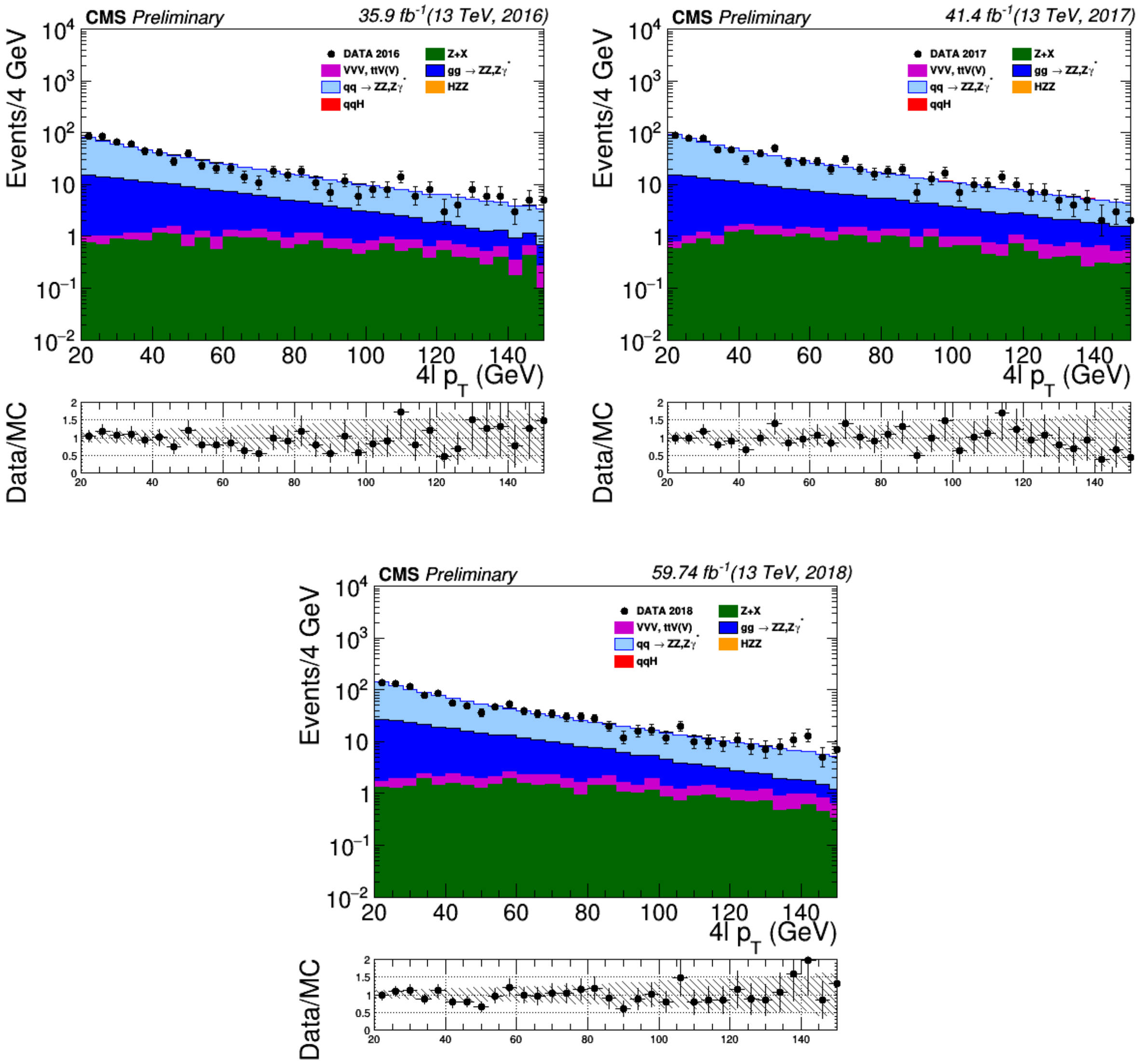


Figure 6.15: Four leptons transverse momentum in the sideband control region for 2016 datasets (top left), 2017 (top right) and 2018 (bottom)

Chapter 7

Extraction of the VBF Higgs signal with machine learning techniques

Machine learning techniques have become increasingly powerful tools for binary classification problems in high energy physics, where the goal is to separate the signal from the background. These algorithms can learn the characteristics of the signal and the background in multidimensional phase space, using complex features and correlations among many variables to best estimate the physical process from which a given event originates. This analysis uses Artificial Neural Networks (ANN) to distinguish the VBF Higgs signal from the background. This chapter describes the general architecture of the ANN used, the learning process, the training strategy, the performance of the trained ANN, and a comparison between data and simulation for the output discriminant.

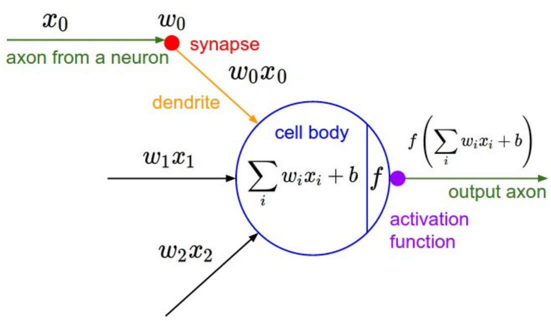
7.1 Artificial Neural Networks

An artificial neural network [153, 154] is a set of assembly interconnected processing units, denoted as nodes or neurons. Each of these neurons performs a specific operation and communicates its output to each other. Such a design is inspired by the brain analogy, which can be thought of as a nonlinear, parallel, and highly complex computer.

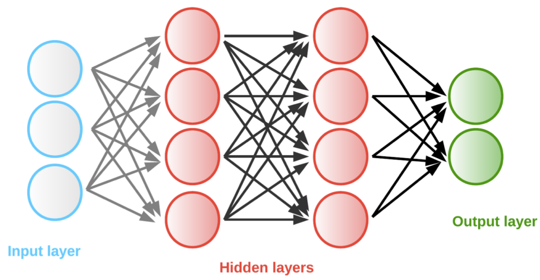
ANN can learn from patterns and generalize the modeling of these patterns. Therefore, it possesses the ability to identify a pattern that the ANN has never seen during the learning process. The neurons contain a function, called activation function, that can receive multiple inputs (arguments), process them, and give a single output value (Fig. 7.1a). Several types of functions are available, and selecting the best function relies on the problem's nature. Different architectures can be used for assembling the neurons. A typical ANN architecture can be seen in Fig. 7.1b. This type of ANN is called Deep Neural Network (DNN) due to its multiple layers. Such an ANN has demonstrated the ability to deal with very complex tasks.

7.1.1 The Learning Process

Machine learning algorithms are usually classified as unsupervised or supervised learning algorithms. This distinction has practical consequences for what concerns their training time; however, both



(a)



(b)

Figure 7.1: (a) Graphical illustration of an artificial neuron used in ANNs. Each input x_i (coming from the hidden neurons or provided by the user) is weighted by w_i and added to a bias b. (b) Graphical illustration of a DNN with two hidden layers.

kinds of algorithms are conceptually similar. Unsupervised learning algorithms only experience "features" but no supervision signal, such as a target or label. This kind of learning can be used in case it is not possible to label the objects (signal/background events) in the dataset. Nevertheless, the ANN can still be trained to find out the likely patterns available in the dataset.

On the other hand, in supervised learning, the dataset is known to contain patterns, and these patterns are known. The algorithm is often viewed as a function that maps the inputs to the outputs. The algorithm consists in optimizing the function based on a criterion and finding the best parameters. The labels can be selected arbitrarily. For instance, one can use 0 or 1 for the labels of the examples for the classification tasks with two classes. Then, a function called *loss* (or cost/risk) can be used in order to verify the performance of labels assignment. This function measures the error between the ANN response and the expected result [154]. The output of the *loss* function is the primary input used to modify the ANN parameters in order to develop the modeling of the data furtherly. The adaption of the ANN parameters (Fig. 7.1a) is done by an algorithm called *back-propagation*, which basically consists of an application of the chain rule for partial derivatives.

Exemplifying such a process, let us consider a neuron j in one of the hidden layers of the ANN. The weight assigned by that neuron j to an input $y_i(n)$ at a given iteration n is denoted by w_{ji}; then such neuron creates a *local field* presented by

$$v_j(n) = \sum_{i=1}^{m} w_{ji}(n) y_i(n) + b_j(n) \tag{7.1}$$

where m represents the total number of inputs received by the neuron, and $b_j(n)$ is the so-called bias. Then, the *local field* is passed to the activation function as an argument (described by f in Fig. 7.1) that produces the *signal function* $y_j(n)$ of the neuron j at iteration n,

$$y_j(n) = f[v_j(n)] \tag{7.2}$$

It is possible to use the *signal error* $e_j(n)$ of the neuron j to compute its instantaneous error (Eq. 7.3), defined as the difference between the neuron *signal function* $y_j(n)$ (the activation function output) and the expected *response* $d_j(n)$. The expected response $d_j(n)$ can be the label identifying a training example for neurons in the output layer or an actual response that a hidden neuron in the intermediate layers should have, based on the example label. Eq. 7.3 is the core of the so-called *loss* or *empirical risk*, and it is an example of *loss* function used here to exemplify the ANN learning process, but there are other types, too.

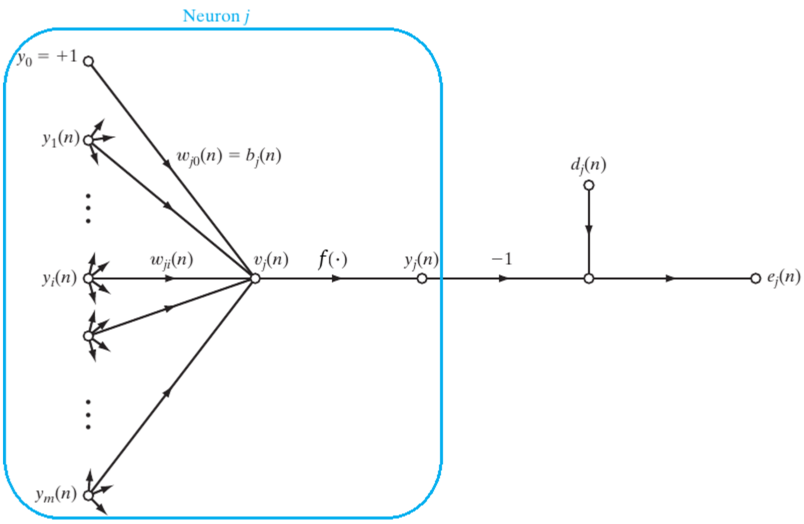


Figure 7.2: The signal flow through the neuron j up to its instantaneous error e_j. [154]

To summarize, the processes described so far can be graphically represented, as shown in Fig. 7.2.

$$\mathcal{E}(n) = \frac{1}{2}e_j^2(n), \quad \text{with} \quad e_j(n) = d_j(n) - y_j(n) \tag{7.3}$$

Now, taking the partial derivative of the *loss*, by applying the chain rule one gets,

$$\frac{\partial \mathcal{E}(n)}{\partial w_{ji}} = \frac{\partial \mathcal{E}(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} \frac{\partial v_j(n)}{\partial w_{ji}(n)}. \tag{7.4}$$

and replacing the derivatives accordingly,

$$\frac{\partial \mathcal{E}(n)}{\partial e_j(n)} = e_j(n) \tag{7.5}$$

$$\frac{\partial e_j(n)}{\partial y_j(n)} = -1 \tag{7.6}$$

$$\frac{\partial y_j(n)}{\partial v_j(n)} = f'[v_j(n)] \tag{7.7}$$

$$\frac{\partial v_j(n)}{\partial w_{ji}(n)} = y_i(n) \tag{7.8}$$

one arrives to

$$\frac{\partial \mathcal{E}(n)}{\partial w_{ji}} = -e_j(n)\, f'[v_j(n)]\, y_i(n). \tag{7.9}$$

Finally, the correction applied to the weights w_{ji} is given by

$$\Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial w_{ji}} \tag{7.10}$$

where the negative sign accounts for the descending gradient in the weights space (that is, it seeks a direction in which the weight change leads to decreasing $\mathcal{E}(n)$), where the η term is the so-called *learning rate* of the back-propagation algorithm. Eq. 7.10 is the core of the synaptics (w_{ji}) update within the neural network. Substituting by Eq. 7.9 into Eq. 7.10 yields

$$\Delta w_{ji}(n) = \eta\ \delta_j(n)\ y_i(n), \quad \text{with} \quad \delta_j(n) = \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} \tag{7.11}$$

The term $\delta_j(n)$ is the *local gradient*. From this point on, there are two situations:

1. The neuron is located in the output layer: in such a case, that neuron is supplied by the labels classifying the examples (which is its *expected response* d_j) and thus, by Eq. 7.3, one has the error and can compute the *local gradient* $\delta_j(n)$;

2. The neuron is in a hidden layer, some complications arise but still Eq. 7.10 holds. Fig. 7.3 shows a schematic of the signal flow in this case. According to Eq. 7.11 and Eq. 7.6 it is possible to redefine the *local gradient* for a hidden neuron j as

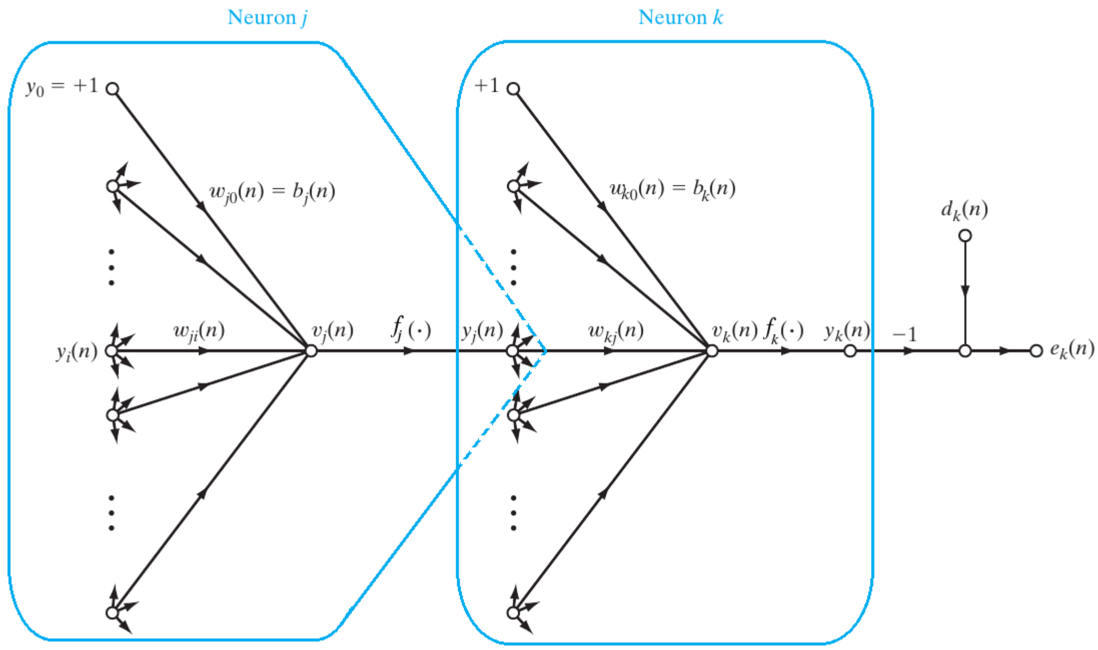


Figure 7.3: The flow of signal through a hidden neuron j up to its *signal function* $y_j(n)$ which is forwarded as an input to an neuron k in an output layer. Note that, the *activation function* can now be different between the layers and they are represented as f_j and f_k for the neurons j and k, respectively. [154]

$$\delta_j(n) = -\frac{\partial \mathcal{E}(n)}{\partial y_j(n)}\frac{\partial y_j(n)}{\partial v_j(n)} = -\frac{\partial \mathcal{E}(n)}{\partial y_j(n)}\ f'[v_j(n)]. \tag{7.12}$$

From Eq. 7.3 and Fig. 7.3 it is clear that $\mathcal{E}(n) = (1/2)\sum_{(k\ \epsilon\ C)} e_k^2(n)$. Differentiating this expression with respect to $y_j(n)$ one has

$$\frac{\partial \mathcal{E}(n)}{\partial y_j(n)} = \sum_{k\ \epsilon\ C} e_k(n)\ \frac{\partial e_k(n)}{\partial y_j(n)} \equiv \sum_{k\ \epsilon\ C} e_k(n)\ \frac{\partial e_k(n)}{\partial v_k(n)}\ \frac{\partial v_k(n)}{\partial y_j(n)} \tag{7.13}$$

Now, from Fig. 7.3 and the previous discussions, one sees that,

$$\begin{aligned} e_k(n) &= d_k(n) - y_k(n) \equiv d_k(n) - f_k'[v_k(n)] && (7.14)\\ \frac{\partial e_k(n)}{\partial v_k(n)} &= -f_k'[v_k(n)] && (7.15)\\ v_k(n) &= \sum_{j=1}^{m} w_{kj}(n) y_j(n) + b_k(n) && (7.16)\\ \frac{\partial v_k(n)}{\partial y_i(n)} &= w_{kj}(n) && (7.17) \end{aligned}$$

and, using Eq(s). 7.15 and 7.17 into 7.13 one has

$$\begin{aligned} \frac{\partial \mathcal{E}(n)}{\partial y_j(n)} &= -\sum_{k\ \epsilon\ C} e_k(n)\ f_k'[v_k(n)]\ w_{kj}(n) \\ &= -\sum_{k\ \epsilon\ C} \delta_k(n)\ w_{kj}(n) && (7.18) \end{aligned}$$

where the definition of the *local gradient* has been used. Finally, using Eq. 7.18 in 7.12, one finds that the *back-propagation* formula for the hidden neuron j is

$$\delta_j(n) = f_j'[v_j(n)] \sum_{k\ \epsilon\ C} \delta_k(n)\ w_{kj}, \quad \text{neuron } j \text{ is hidden.} \tag{7.19}$$

Note that the sum in Eq. 7.19 is over all the neurons in the output layer. Also, note that $e_k(n)$ can be extended to internal layers, such that it will correspond to the signal errors coming from neurons in the forward hidden layers.

7.2 Strategy for Developing and Training the Artificial Neural Networks

In this analysis, the base software used to build up and train the NNs is Keras [155]. Keras is a widely used and powerful set of python libraries that can quickly create any known NN. It contains sets of modules that simplifies much of the coding needed by the user.

7.2.1 Preparation of the Datasets

The neural networks were trained with the simulated samples described in Section 5.2, using only with the events satisfying the selection criteria of VBF-SR. In order to optimize the analysis and adequately handle the usage of the third jet in the events, two jet-based categories were defined, henceforth referred to as "*Njet2*" and "*Njet3*". The category *Njet2* includes the events with exactly two jets, while the category *Njet3* consists of the events with at least three jets. For each category, a ANN has been developed.

In order to ensure an adequate distribution of events after splitting to form the exclusive training and testing sets, the following procedure was adopted for the VBF-SR events in each jet-based category:

1. Each simulated process contains three samples, one for each final state (4μ, $4e$, and $2e2\mu$). These three samples were merged into a single sample, with events from each final state being randomized, such that the final states randomly populated the merged sample.

2. The merged and randomized samples for each process were divided into two parts, one for the training and validation steps and the other for the testing phase.

The signal and background events used to extract the final results constitute the test dataset. The training dataset is used to train the discriminator and is thoroughly orthogonal to the test dataset. Whenever a discriminator is trained, an additional dataset is needed to verify its expected performance; this step is called validation. The validation dataset must be independent of both the training and test datasets. The validation step is mandatory to verify the absence of overtraining, compare discriminators built with different hyper-parameters, and hence optimize the results. Furthermore, processes with too few events were not included in the training or testing sets. As explained in Section 7.1 the ANN training is very similar to a fit procedure, and therefore few examples (points) produce fits with large errors.

Two independent training were performed: one for 2017/2018 data and another for 2016 data in order to account for the upgrade done to the detector between those two periods.

7.2.2 The Training Procedure

Training with MVA methods is a procedure that needs to be done on multiple fronts. It is difficult to guarantee that a set of MVA training parameters is the optimal choice for a set of inputs since there are an infinite number of combinations that one can build. These configurations (including the training dataset) can significantly influence the development of the MVA training. In order to optimize the study, thousands of ANN architectures were developed, and multiple parallel scans were performed. These scans can run over different input sets and different parameters configured for training. The result of each ANN training can be retrieved to produce plots that are used to validate and classify the quality of each training (e.g., one can check if overfitting has occurred). The set of tested inputs includes combinations of the kinematic variables for leptons and jets (p_T, η and ϕ), as well as the quark-gluon likelihood for jets. The scanned ANN parameters are summarized in the Table 7.1. These parameters are:

Topology

Topology represents the architecture of the ANN: the number of hidden layers, number of neurons in each layer, and neuron type. Several ANN architectures are tested, including simple networks consisting of one layer of neurons and deep networks. The activation function of the hidden layer is optionally nonlinear. Most neural networks nowadays use the *Rectified Linear Unit* (ReLU) activation function, which sets the negative values to zero and leaves the positive ones unchanged. Several functions are also used that are similar to ReLU but avoid zero values.

Functions representable with DNNs, with ReLU or similar activation functions, can separate a number of regions that is exponential in the depth of the network [156]. In the case of one hidden layer, an exponential number of hidden units is required. Therefore, choosing a model with more than one hidden layer may be advantageous: the number of nodes required for optimal capacity is significantly reduced with a deep model. In the present analysis, a DNN provided the most beneficial topology.

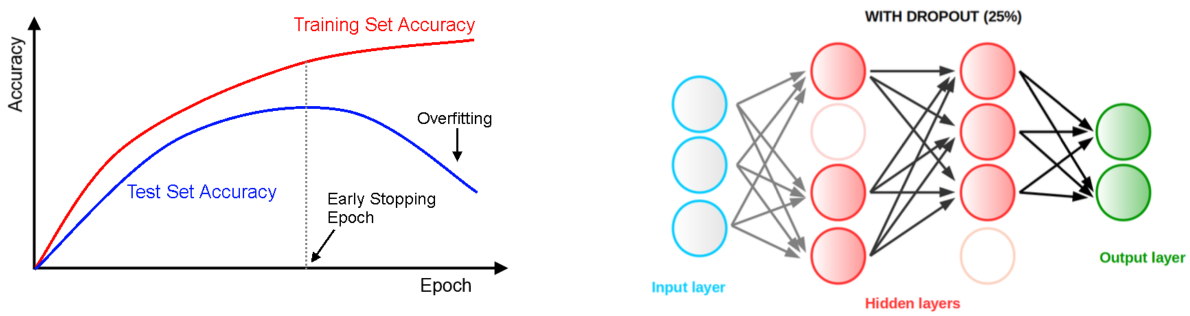
Optimization

The pure optimization of a model would be finding the global minimum of the cost function using gradient descent or other techniques. However, this is not possible for ANNs and DNNs due to the nonconvexity of the loss functions. The goal of model optimization is to achieve satisfactory performance according to the performance metrics chosen for the task. The loss function itself is a proxy of the performance metrics, which is minimized as they improve. Training can be terminated if the performance metric no longer improves. There are several options in Keras (SGD, Adam, AdaDelta, RMSprop, etc.), and after testing most of them, it was found that Adam [157] and AdaDelta [158] give a satisfactory performance.

Regularization

Regularization is any modification that can be made to a learning algorithm to reduce its generalization error. Among the techniques that help in regularization is the so-called "early stopping", a parameter that sets the number of epochs Keras should wait when no improvement in the loss function is observed. If after this number of epochs still no improvement is seen, the training is stopped. Representation of the use of early stopping in ANN training is shown in Fig. 7.4a, which illustrates how using early stopping in the ANN training reduces and puts under control overfitting.

Another method adds a typical DNN regularization called "dropout". Dropout [159] provides a computationally inexpensive but powerful method for regularizing DNNs. When training with dropout, a fraction of nodes in a layer is randomly zeroed at each iteration (see Fig. 7.4b). Dropout trains the ensemble consisting of all subnets formed by removing non-output units from an underlying base net. Adding dropout to a DNN ensures regularization and makes the network more robust.



(a) (b)

Figure 7.4: A visualization of the methods used to regularize NN training is shown. (a) Illustration of the use of early stopping in NN training. (b) Illustration of a DNN with two hidden layers, using a dropout rate of 25%.

Batch size

It represents the number of events in a subset from the training set used to compute the gradients and update the w_i's and b in each neuron.

Epochs

It indicates the number of iterations over the entire training set. The total number of iterations is a combination of the size of the training set, the batch size and the number of epochs. For instance, setting training of 10 epochs and a batch size of 10 for a training set size of 100 means that Keras performs 10^2 updates on the w_i's and b's.

Parameter	Tested options
Inputs	leptons/jets(p_T,η,ϕ), jets Q/G L
Dropout rate	10, 20, 30, 40%
Topologies	30:20:10, 21:13:8, 10:10:10:10, 50:50:50:50,....
Early stop	50, 100, 600, 3000
Minimizer	SGD, Adam, Adagrad, Adadelta, RMSprop
Batch size	5, 32, 64, 128, 786
Neuron	ReLU, SeLU
Loss Scaling	XS (process total XS), $\sigma.\epsilon.BR$ (event weight)

Table 7.1: Summary of the scanned parameters during NN training.

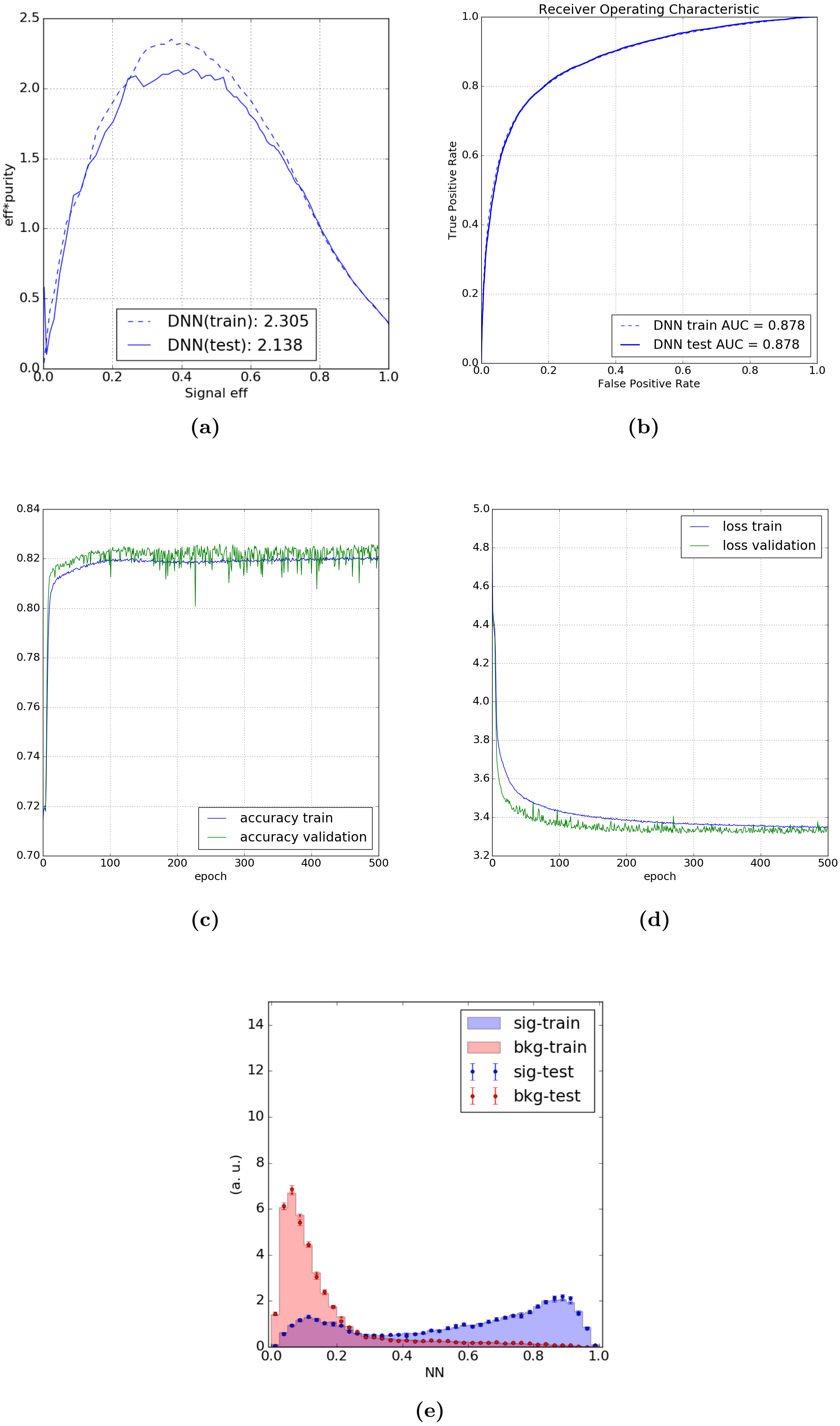
7.2.3 The Training Performance

The ANNs obtained through the procedures explained in the previous Section are produced together with some validation plots that validate the choice of the best ANN architecture, Figs. 7.5 - 7.8 show these plots.

These plots include the metrics adopted ROCs curves and efficiency times purity ($\epsilon_s.\pi$). In addition, the plots which help to check the existence of the over-fitting for a given configuration: the NN distributions for the training and test sets, accuracy, and loss. The ROC curves[1] from training and test sets can also be used to check the existence of the over-fitting. Then finally, the selected ANN is chosen to be the ones with the highest ($\epsilon_s.\pi$) for each jet-based category.

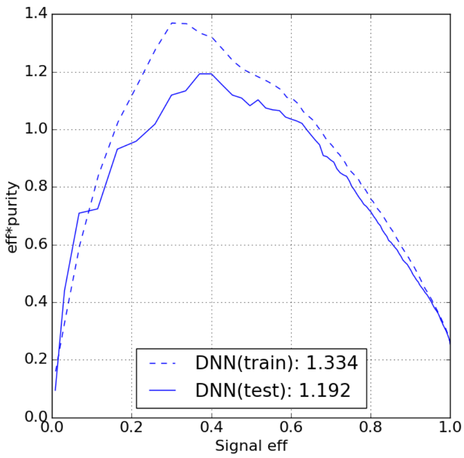
In order to understand if the data well describe the ANN response, the ANN discriminant is plotted in the four leptons sideband control region $m_{4l} < 118$ or $m_{4l} > 130\ GeV$, for events with at least two jets, as shown in Fig. 7.9. The plots show a good agreement between data and MC.

[1]The ROC curve is a graph showing the classification model performance at all the classification thresholds.

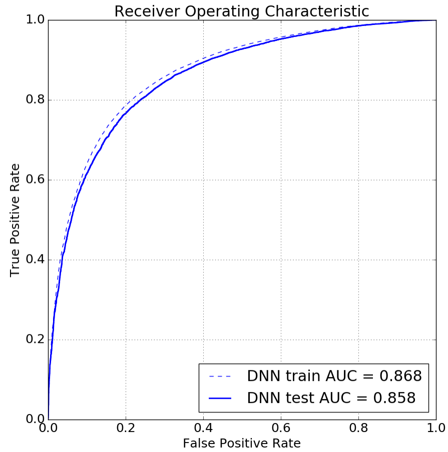


(a) (b) (c) (d) (e)

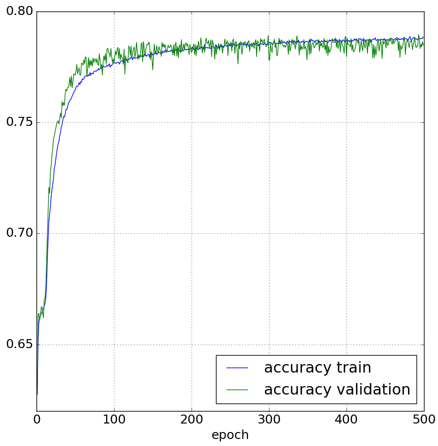
Figure 7.5: Validation plots produced during ANNs training for Njet2 category for 2018 and 2017. (a) $\epsilon_s.\pi$ vs ϵ, (b) ROC curves computed for both training and testing sets, using all backgrounds together, (c) Accuracy, (d) Loss, and (e) ANN distributions.



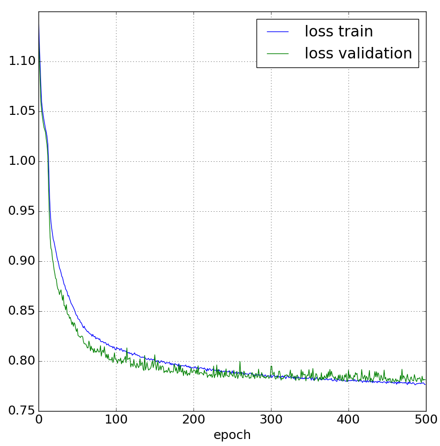
(a)



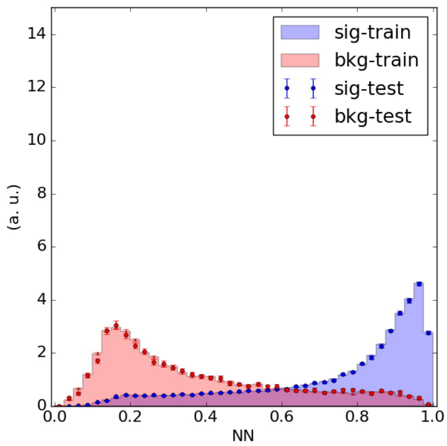
(b)



(c)



(d)



(e)

Figure 7.6: Validation plots produced during ANNs training for Njet3 category for 2018 and 2017. (a) $\epsilon_s.\pi$ vs ϵ, (b) ROC curves computed for both training and testing sets, using all backgrounds together, (c) Accuracy, (d) Loss, and (e) ANN distributions.

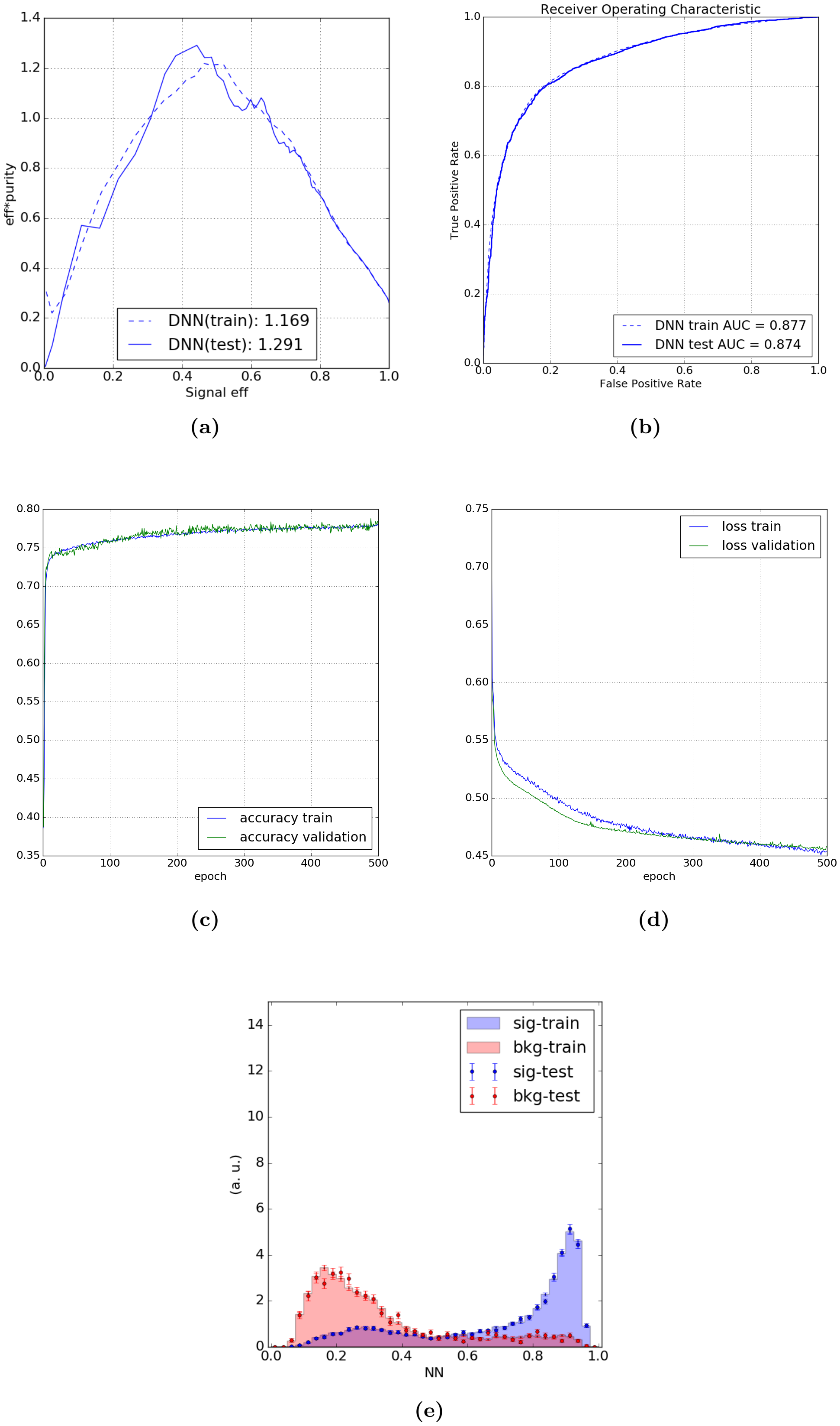
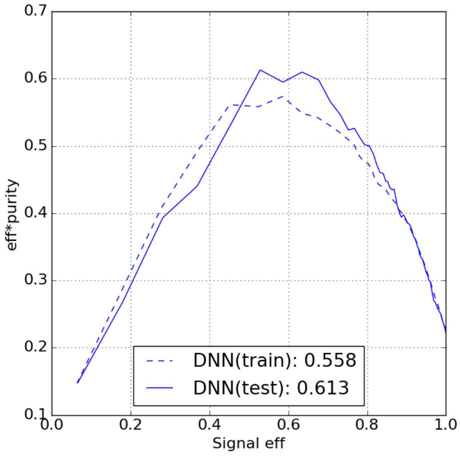
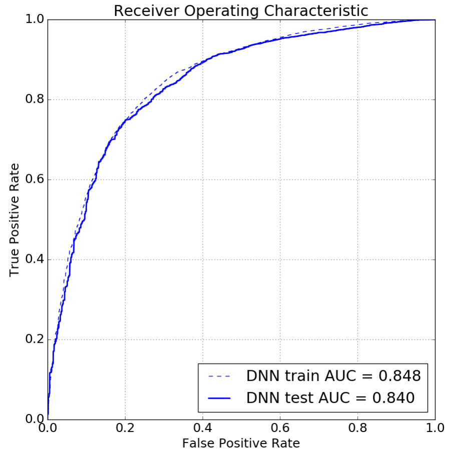


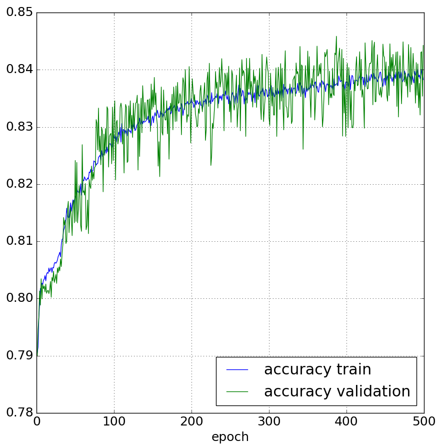
Figure 7.7: Validation plots produced during ANNs training for Njet2 category for 2016. (a) $\epsilon_s.\pi$ vs ϵ, (b) ROC curves computed for both training and testing sets, using all backgrounds together, (c) Accuracy, (d) Loss, and (e) ANN distributions.



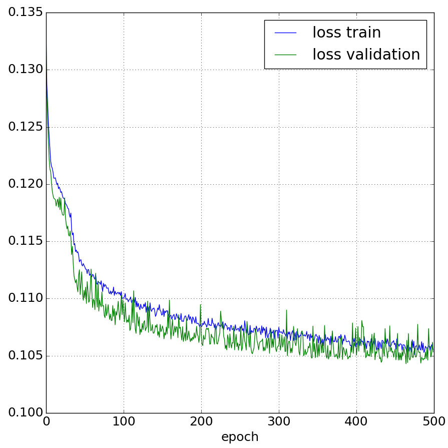
(a)



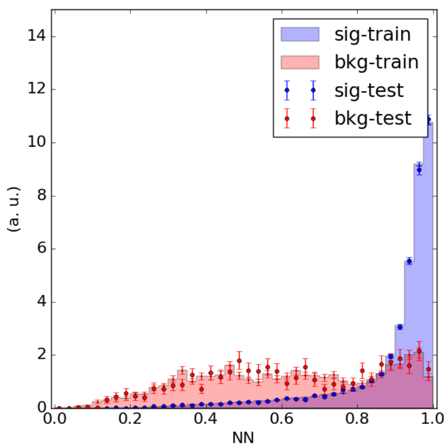
(b)



(c)



(d)



(e)

Figure 7.8: Validation plots produced during ANNs training for Njet3 category for 2016. (a) $\epsilon_s.\pi$ vs ϵ, (b) ROC curves computed for both training and testing sets, using all backgrounds together, (d) Loss, and (e) ANN distributions.

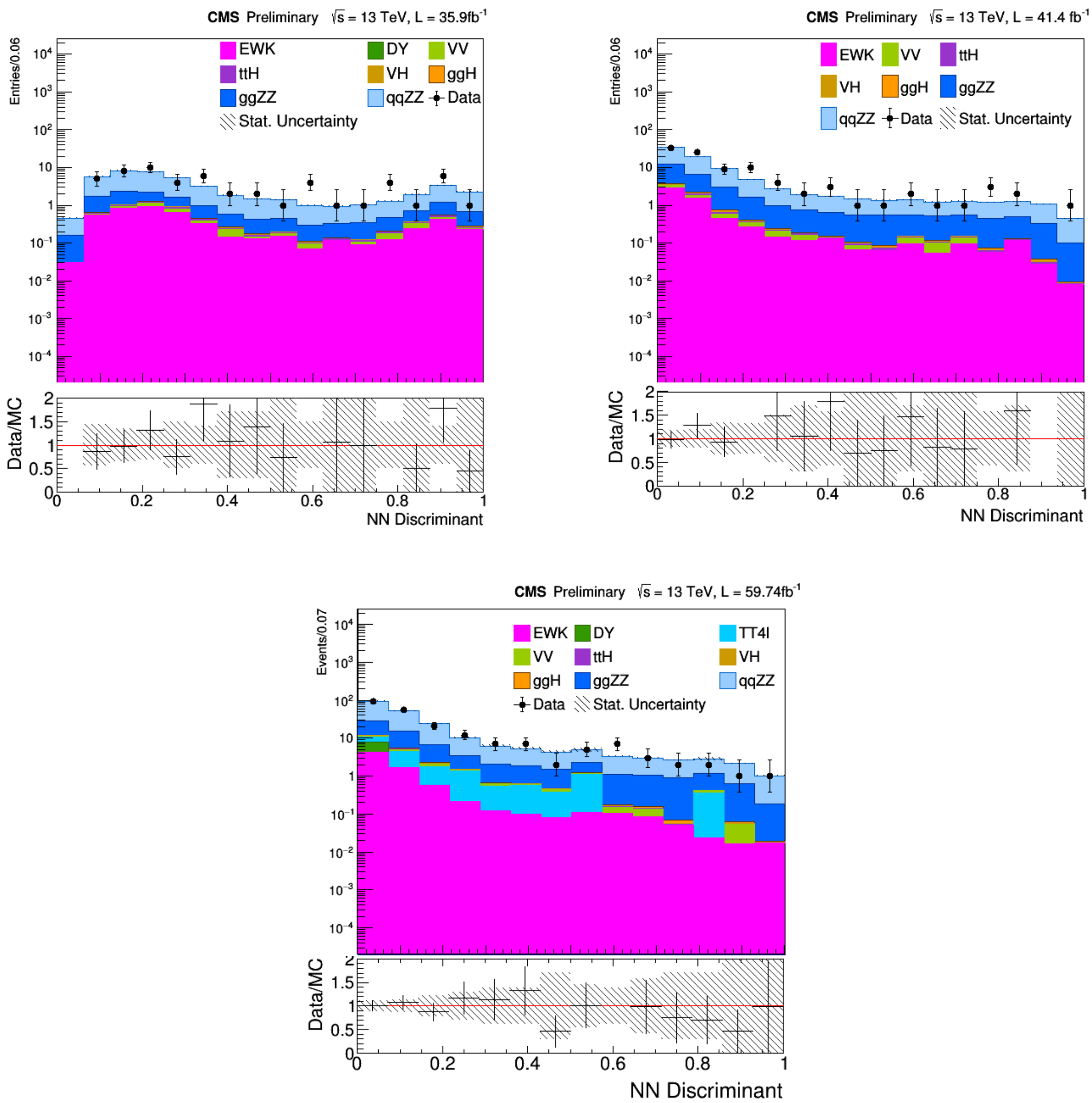


Figure 7.9: ANN discriminant in the sideband control region for 2016 (top left), 2017 (top right) and 2018 (bottom) datasets.

Chapter 8

Results and statistical interpretations

This chapter presents the results of the search for VBF Higgs production conducted in this analysis. The various sources of systematic uncertainties that affect the analysis are discussed, their impact is evaluated, and finally, the results and interpretations are shown.

8.1 Systematic uncertainties

Systematic uncertainties that affect this analysis and their treatment are debated in this section. Measurement uncertainty has two components: statistical and systematic. Statistical uncertainty depends on the amount of data collected, while systematic uncertainty results from the inherent imprecision of the method of measurement used. The systematic uncertainties are divided into theoretical and experimental uncertainties. Theoretical uncertainties are the sources of error that arise from the modelling of the signal and background. These can be of experimental origin, related, for instance, to insufficient knowledge of the detector response. In addition, there are uncertainties directly related to the signal and background modelling by the ANN, which is related to the uncertainties on its input objects. The systematic uncertainties can affect both the final significance and the signal strength uncertainty. Systematic uncertainties are implemented as normalization variations of a process or shape variations of the fitted distribution. If a shape uncertainty is used, as in the present case, the uncertainty is either applied to the relevant objects and propagated through the event selection and the DNN computation or treated as an event weight when building the variations about the nominal distribution.

8.1.1 Theoretical uncertainties

Theoretical uncertainties are used as a normalization uncertainty, affecting both signal and background estimation. Those uncertainties are summarized in Table 8.1, and described in the following:

- **Parton distribution function (PDF)**: the uncertainty from the choice of the PDF set is determined by calculating the root mean square of the variation when using different replicas of the default PDF set, following the recommendations reported in [160].

- **Renormalization and Factorization**: the uncertainty on the renormalization and factorization scales is estimated by changing these scales between 0.5 and 2 times their nominal value while keeping their ratio between 0.5 and 2 [161]

- **Strong coupling**: is the uncertainty that originate from the choice of the strong coupling (α_s), described in [131].

- **K-factor**: the uncertainty of the K factor used for the gg → ZZ and qq → ZZ samples is applied as described in [122].

- **Branching ratio**: The uncertainty on the branching ratio of the H → 4l is applied to the yields of all Higgs processes [122].

Source	Magnitude (%)
PDF set and α_s VBFH	2.1
QCD scale VBFH	0.3 - 0.4
PDF set ggH	1.8
α_s ggH	2.59 - 2.62
QCD scale ggH	4.27 - 6.49
PDF set and α_s WH	1.3
QCD scale WH	0.5
PDF set and α_s ZH	1.6
QCD scale ZH	2.7 - 3.5
PDF set and α_s ttH	3.6
QCD scale ttH	6.0 - 9.2
PDF set and α_s bbH	3.2
QCD scale bbH	4.6 - 6.7
PDF set and α_s ($q\bar{q} \to ZZ$)	3.1 - 3.4
QCD scale ($q\bar{q} \to ZZ$)	3.2 - 4.2
Electroweak correction ($q\bar{q} \to ZZ$)	0.1
PDF set and α_s ggZZ	3.2
QCD scale ggZZ	4.6 - 6.7
Electroweak correction ggZZ	10.0
BR($H \to ZZ \to 4l$)	2.0

Table 8.1: Summary of theoretical systematic uncertainties.

8.1.2 Experimental Uncertainties

- **Luminosity**: The rate of uncertainty on the integrated luminosity amounts to 2.6% for the data recorded in 2016 [119] and 2017 [120], and to 2.3% for the data recorded in 2018 [121]. This uncertainty arises from the stability of the detector response during data taking, affecting both signal and background. This uncertainty is used as a normalization uncertainty on the event yield.

- **Lepton identification and isolation efficiency**: This uncertainty is related to electron and muon reconstruction and selection efficiency, including triggers, the significance of impact parameters, identification and isolation requirements. This uncertainty is obtained by varying the corresponding scale factors within their uncertainties obtained from the tag-and-probe measurement in Z-boson resonance events in data and MC. While we rely on low-mass di-muon resonances to measure muon efficiency in the low p_T regions, the measurement of electron efficiency relies solely on the Z-boson resonance, leading to a larger uncertainty in the low

p_T region. These uncertainties amount to 11-15.5% for electrons and 1-2.3% for muons [37], affecting both the shape and the rates of the signals and backgrounds.

- **Reducible background estimation**: Uncertainties in the Z+X estimation, described in Section 6.2.2.4, varies between 30% and 41%, depending on the final state. This uncertainty is used as a normalization uncertainty on the event yield for the Z+X background.
- **Lepton momentum scale**: The uncertainties in lepton momentum measurements are varied up and down within one standard deviation and propagated through the DNN computation, resulting in a shape uncertainty.
- **Jet energy corrections (JES/JER)**: The uncertainties associated with the jet energy scale (JES) are obtained by shifting the energy scale applied to the reconstructed jets up and down, within one standard deviation around the central value and propagated through the DNN, resulting in a shape and normalization uncertainty. The uncertainty propagation for jet energy resolution (JER) is analogous to the jet energy scale.
- **Quark-Gluon likelihood**: The polynomial corrections described in section 3 are applied to the MC samples leading to the nominal distributions. Then, the variations to the nominal distributions are computed by applying those corrections twice for up variations and not applying the corrections for the down variations, and the quark-gluon systematic variations propagated through the DNN, resulting in a shape uncertainty.

8.1.2.1 Systematic Uncertainty on the 3^{rd} Jet

In order to properly handle the systematic uncertainty associated with the usage of the third jet, two jet-based categories were defined in this analysis, as mentioned in Section 7.2.1. This uncertainty arises from the fact that the third jet is generated at NLO in the VBF-H3J sample (see Section 5.2.1) while at LO in the background.

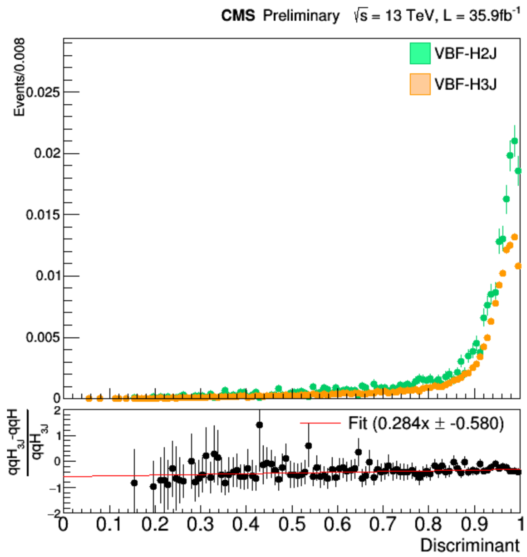
A procedure for estimating the uncertainty on the 3^{rd} jet affecting the ANN distribution has been estimated as follows:

- The events in the *Njets3* category were fed into the ANN from each sample (VBF-H2J and VBF-H3J), and the uncertainty in both normalization and shape variation between the two processes has been evaluated from the ratio between their shapes. This procedure was performed for each four-lepton final state and separately for each year. The ANN distributions of the two samples are shown in Figs. 8.1 - 8.3 with their ratio.
- In order to smooth the statistical variations bin by bin, which arises mainly in the VBF-H2J with fewer events, a fit is performed to the ratio. Then, the up and down uncertainties are introduced around the distribution of the VBF-H3J sample, which is the nominal one, using the linear function derived from the ratio. Such uncertainty is introduced in the statistical analysis following the procedure for shape analysis, as explained in Section 8.2.

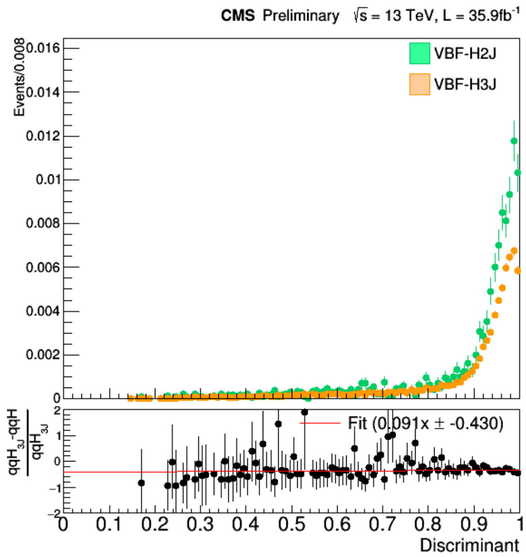
The summary of experimental systematic uncertainties is reported in Table 8.2.

type	2016	2017	2018
Luminosity	2.6%	2.3%	2.5%
Leptons ID and reco eff	1.6 – 15.5%	1.1 – 12.1%	1.0 – 11%
Z+X	30 – 41%	30 – 38%	30 – 37%
Lepton p_T	shape	shape	shape
Jet energy scale	shape	shape	shape
Jet energy resolution	shape	shape	shape
Q/G L	shape	shape	shape
3^{rd} jet	shape	shape	shape

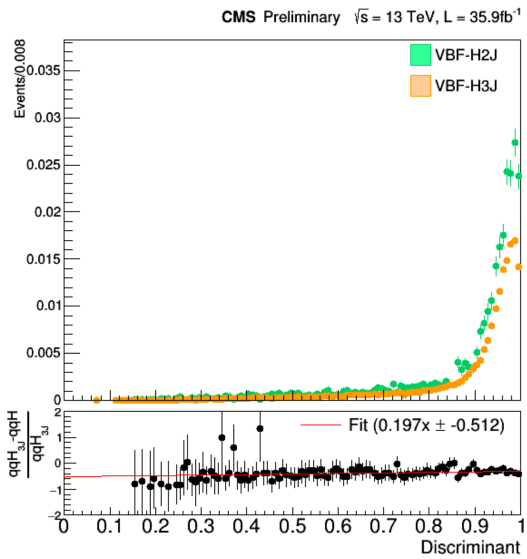
Table 8.2: Summary of experimental systematic uncertainties.



(a)

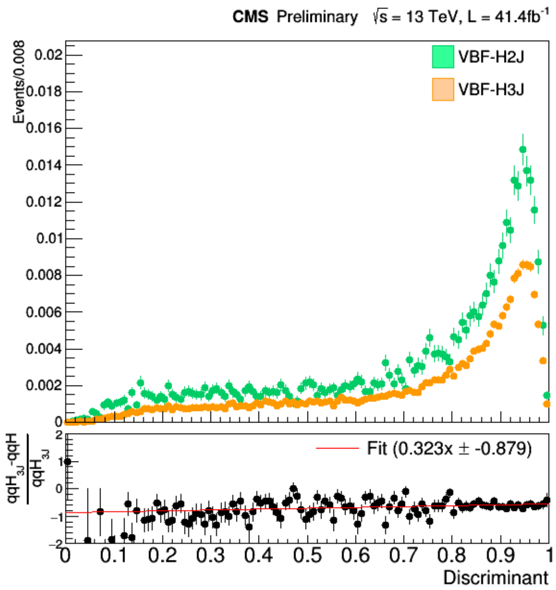


(b)

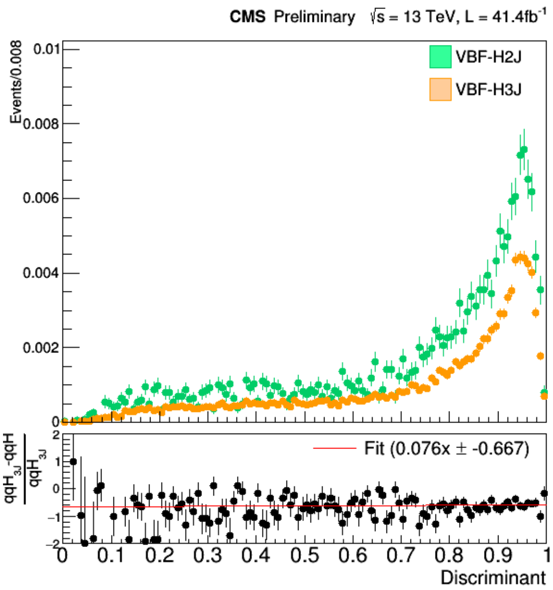


(c)

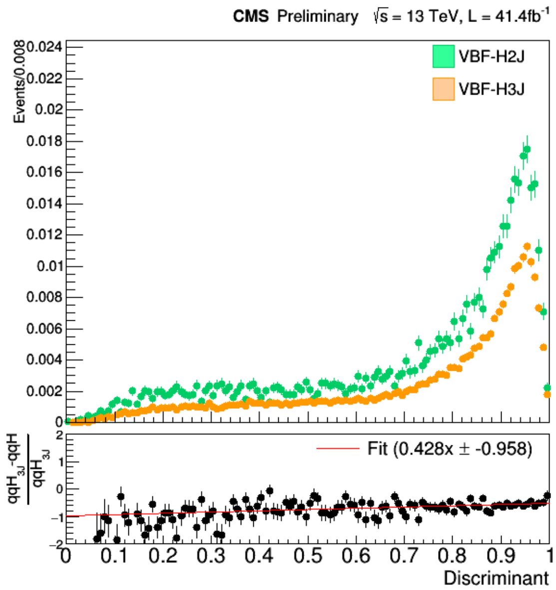
Figure 8.1: ANN distributions in the *Njets3* category from VBF-H2J and VBF-H3J and their ratio for (a) 4μ, (b) $4e$ and (c) $2e2\mu$ for 2016 dataset.



(a)

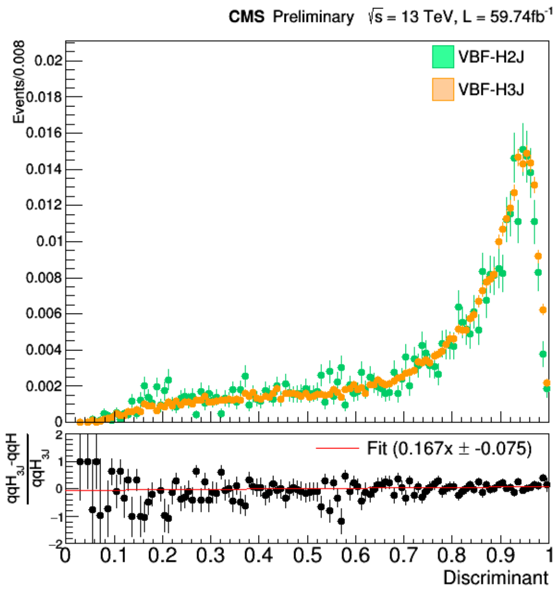


(b)

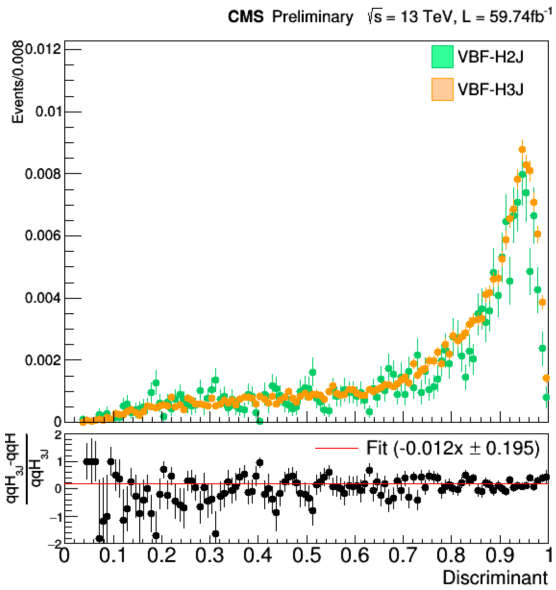


(c)

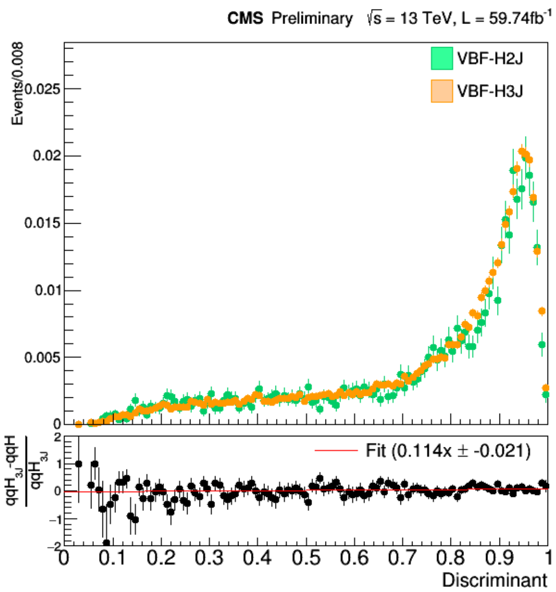
Figure 8.2: ANN distributions in the *Njets3* category from VBF-H2J and VBF-H3J and their ratio for (a) 4μ, (b) $4e$ and (c) $2e2\mu$ for 2017 dataset.



(a)



(b)



(c)

Figure 8.3: ANN distributions in the *Njets3* category from VBF-H2J and VBF-H3J and their ratio for (a) 4μ, (b) $4e$ and (c) $2e2\mu$ for 2018 dataset.

8.1.3 Discriminants Shape Uncertainties

The systematic uncertainties on the shape of the ANN output were estimated by studying the impact of the input variables' systematic uncertainties. The main types of systematic uncertainties in data classification are the ones that change the ANN output value that can cause event migration across the regions defined by the discriminant for signal and background. Therefore, one should consider that issue when applying specific discriminant cuts. This effect is exacerbated in the shape analysis as bins shift up and down due to the uncertainties in the binned analysis. In addition, there is an impact from the bin width: a smaller bin width increases the probability of event migration across the bins, and more significant fluctuations than in the case with larger bin width are expected.

In order to estimate the uncertainties propagated through the ANNs a procedure based on the one adopted by [162–164] is used. In such a procedure, the systematic uncertainties of the objects used as inputs for the ANNs were applied to the nominal input value. Those are the uncertainty in the leptons' momentum, the jet energy scale, jet energy resolution, and quark-gluon Likelihood.

The ANNs were fed with the nominal values of the inputs, without applying any systematic uncertainty shifts, and with $\pm 1\sigma$ shifts from the nominal values. These shifts are produced for each variable, where there are $N_{(Inputs)} \times [1 + 2 \times N_{(InputUncertainties)}]$ output values for each event after all shifts have been done, and consequently the same number for the ANN distributions. Fig. 8.4 shows the impact of the systematic uncertainties on the shape of one of the trained ANN for the signal (qqH) and the dominant background (ggH).

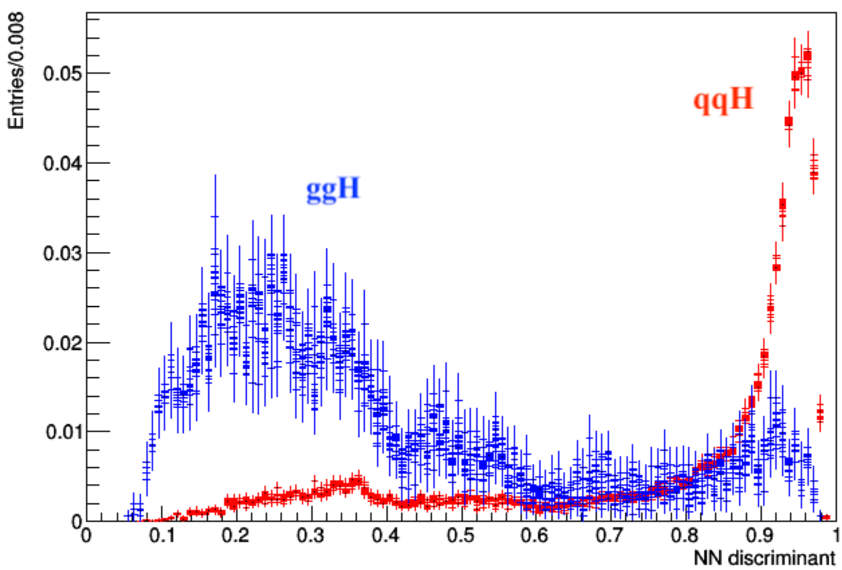


Figure 8.4: Example of the impact of the systematic uncertainties on the shape of a ANN. The "qqH" signal is shown in red, and the dominant background "ggH" is shown in blue; the plot includes events from the 4μ channel of the 2016 dataset.

8.2 Statistical treatement

The statistical procedure to evaluate the signal presence or absence consists in interpreting the observed data compatibility with the background-only hypothesis (H_B) or with the signal plus background hypothesis (H_{S+B}). The estimation is done based on the binned distributions of suitably chosen discriminating observables; the ANN distributions are used for this purpose in the present analysis.

8.2.1 Maximum likelihood

In each bin i of the distributions of the discriminating ANN, considered in the analysis, the expected yields under the H_{S+B} hypothesis can be parameterized as:

$$\nu_i(\mu_{qqH}, \vec{\theta}) = \mu_{qqH} \cdot s_i^{qqH}(\vec{\theta}) + b_i(\vec{\theta}) \tag{8.1}$$

where s_i^{qqH} and b_i are the expected yields of the VBF Higgs signal and the background in the bin i, respectively. In the context of the Higgs boson searches, the signal yields are fixed to their SM prediction; they are scaled by the signal strength modifiers

$$\mu_{qqH} = \frac{\sigma_{qqH}}{\sigma_{qqH}^{SM}} \tag{8.2}$$

which are the *parameters of interest* (POI) of the model. In the background-only hypothesis H_B, the model considers $\mu = 0$.

The predictions for the signal and background yields in Eq. 8.1 are subject to the systematic uncertainties described in Section 8.1. A *nuisance parameter* θ represents each independent source of systematic uncertainty, collectively denoted with $\vec{\theta}$. The rate uncertainties directly affect the signal and background yields, while the shape uncertainties affect the repartition of these yields across the bins of the discriminating observable.

A binned likelihood function combining all known information regarding the model parameters is constructed to estimate the values of the POIs and their associated uncertainties. This function quantifies the compatibility between the observed data and the prediction for given values of the POIs and the nuisance parameters. It is expressed as the product of the Poisson probabilities to observe n_i events in the bin i, scaled by the nuisance probability function $\rho(\theta|\tilde{\theta})$, as

$$\mathcal{L}(data|\mu, \vec{\theta}) = \prod_i \frac{(\mu s_i(\vec{\theta}) + b_i(\vec{\theta}))^{n_i}}{n_i!} e^{-(\mu s_i(\vec{\theta}) + b_i(\vec{\theta}))} \cdot \rho(\theta|\tilde{\theta}) \tag{8.3}$$

The POIs and nuisances parameters estimation is done via the maximum likelihood method, which finds the point $\mu \equiv \hat{\mu}$ and $\vec{\theta} \equiv \hat{\vec{\theta}}$, where the values $\hat{\mu}$ and $\hat{\vec{\theta}}$ are called the best estimates in the parameter space, making the likelihood function $\mathcal{L}(data|\mu, \vec{\theta})$ maximal, namely the point for which the observed data are the most probable. For computational simplicity, it is common to minimize -2 $ln(\mathcal{L})$ instead, as the maximum likelihood value means the minimum negative log-likelihood. The uncertainty on the best estimate $\hat{\mu}$ is obtained by determining the lower and upper bounds, μ_{min} and μ_{max}, for which the value of -2 $ln(\mathcal{L})$ exceeds the minimum by one unit, corresponding to a coverage probability of 68%. The differences $\delta_+ = \mu_{max} - \hat{\mu}$ and $\delta_- = \hat{\mu} - \mu_{min}$ correspond to the combined uncertainties associated to $\hat{\mu}$.

In the case of the present analysis, for the signal strength estimation, the likelihood ratio is performed with $\mu = 1$, which represent the expected SM VBF signal and $\hat{\mu}$ is left to float during the fit on the observed events, which represent the VBF signal estimated from observed data.

The sources of uncertainty included in this analysis are described in Section 8.1. The shape of their associated nuisance probability function in Eq. 8.3 is modeled by the log-normal distribution, which is defined as

$$\rho(\theta|\tilde{\theta}) = \frac{1}{\sqrt{2\pi}\ln k} \exp\left(-\frac{(\ln(\theta/\tilde{\theta}))^2}{2(\ln k)^2}\right)\frac{1}{\theta} \tag{8.4}$$

where $\tilde{\theta}$ is the initial estimate of the parameter and $k = 1 + \epsilon$, ϵ being the relative scale of the uncertainty.

8.2.2 Hypothesis testing

In order to drive reliable conclusions from the results of the maximized likelihood, a hypothesis test is typically specified in terms of a statistic test. This is defined in such a way as to quantify, within observed data, behaviours that distinguish the H_{S+B} hypothesis from the H_B hypothesis. A standard method used for this purpose is the *profile likelihood ratio* (PLR)

$$\lambda(\mu) = \frac{\mathcal{L}(data|\mu, \hat{\hat{\vec{\theta}}}_\mu)}{\mathcal{L}(data|\hat{\mu}, \hat{\hat{\vec{\theta}}}_\mu)} \tag{8.5}$$

The numerator corresponds to the case where μ is fixed at a specific value, and $\hat{\hat{\vec{\theta}}}_\mu$ is the associated *conditional maximum likelihood estimator*, i.e, the value of $\vec{\theta}$ that maximizes the likelihood for the fixed μ. The denominator corresponds to the maximized likelihood obtained when both μ and $\vec{\theta}$ are left floating freely. By construction, the PLR is bounded between 0 and 1, and higher values indicate better compatibility between the data and the H_{S+B} hypothesis. It is common to use the logarithmic form of PLR, defined as $q_\mu = 2\ln\lambda(\mu)$. When measuring a single parameter μ, q_μ is expressed graphically as scans of q_μ for a given μ within a given interval, with the minimum being the best estimate. The intervals with 68% and 95% confidence levels (CL) are then derived from the conditions $q_\mu < 1$ and $q_\mu < 3.84$, respectively.

8.2.3 Significance

The statistic test q_μ under the background-only hypothesis H_B can be used to estimate the *significance* of an observed result. This is done using the *p-value* p_0, representing the probability that the background-only model reproduces a larger fluctuation than the observed data. It is expressed as

$$p_0 = \int_{q_{obs}}^{\infty} f(q_0|H_B)dq_0 \tag{8.6}$$

where $q_0 = -2\ln\lambda(\mu = 0)$, and $f(q_0|H_B)$ is the probability density function of q_0 under the hypothesis H_B. The hypothesis H_B is excluded if p_0 is below a predefined threshold.

The p-value is typically converted into an equivalent significance Z, which is defined as:

$$Z = \Phi^{-1}(1 - p_0) \tag{8.7}$$

which is expressed in units of Gaussian standard deviations σ, where Φ^{-1} is the inverse of the standard Gaussian cumulative distribution (or *quantile*). For a signal such as the Higgs boson, the particle physics community claims *evidence* when a significance of 3σ is reached ($p_0 = 1.4 \times 10^{-3}$) and an *observation* when a significance of 5σ is reached ($p_0 = 2.9 \times 10^{-7}$).

8.2.4 Expected results

The expected results are instrumental in estimating the sensitivity of the experiment. The standard procedure in CMS estimates the expected results is to develop an analysis in a blinded manner, meaning the strategy is defined on studies based solely on the simulation or data in signal-free regions. In principle, this consists in generating a large set of pseudo-data (toys) distributed according to the likelihood function and extracting the nominal values of the POI from the median. An excellent approximation is provided by the Asimov dataset [165], a single representative dataset in which the estimates of the parameters are set to their expected values, and the statistical fluctuations are suppressed. The usage of the Asimov dataset allows the impact of the systematic uncertainties and the validity of the statistical method to be verified before looking at the data, i.e., unblinding.

8.3 Results

This section summarizes the yields and distributions obtained after the VBF Higgs signal selection. In addition, it presents the validation of the statistical model, including the impact of the systematic uncertainties on the signal strength and the goodness-of-fit test. Finally, the best fit values for the signal strength and the significance of the Higgs VBF production mode are presented.

8.3.1 Pre-fit yields and distributions

Since this analysis uses up to three jets, the selected events in the VBF-SR, were split into two categories according to the number of jets in each event that pass the jet selection criteria. These two categories have a different ANN model and a different set of events for each of them, and hence such categories, labeled as *Njets2* and *Njets3*, are orthogonal. For simplicity, the discriminant from the two categories will be referred to as $D^{NN}_{qqH,2J}$ and $D^{NN}_{qqH,3J}$ from now on.

The event yeilds from signal and background after the VBF-SR selections are reported in Tables. 8.3 - 8.5 for each jet category (*Njets2*, *Njets3*) and for their combination, respectively. The uncertainties shown include both statistical and systematic sources.

Fig(s). 8.5 - 8.8 (a), (b) show the distribution of the selected ANN discriminant in the *Njets2* and *Njets3* categories for each year and for the combined Run2, while Figs. 8.5 - 8.8 (c) show their combination. The three four-lepton final states are combined in these distributions.

Process	4μ	4e	$2e2\mu$	4l
ggH	8.09±0.31	3.79±0.34	9.85±0.60	21.73±0.76
VH	0.75±0.01	0.38±0.02	0.90±0.04	2.03±0.05
ttH	0.02±0.00	0.02±0.00	0.02±0.00	0.05±0.00
qqZZ+ZZJJ	1.84±0.06	0.83±0.07	2.15±0.13	4.83±0.16
ggZZ	0.31±0.02	0.15±0.02	0.19±0.02	0.66±0.04
Z+X	0.30±0.04	0.17±0.04	0.47±0.06	0.94±0.09
$\sum$ backgrounds	11.32±0.32	5.34±0.36	13.57±0.62	30.23±0.78
qqH (signal $m_H = 125GeV$)	2.57±0.25	1.28±0.23	3.15±0.43	7.00±0.55
Total expected	13.89±0.41	6.62±0.42	16.72±0.75	37.23±0.96
Observed	18	8	15	41

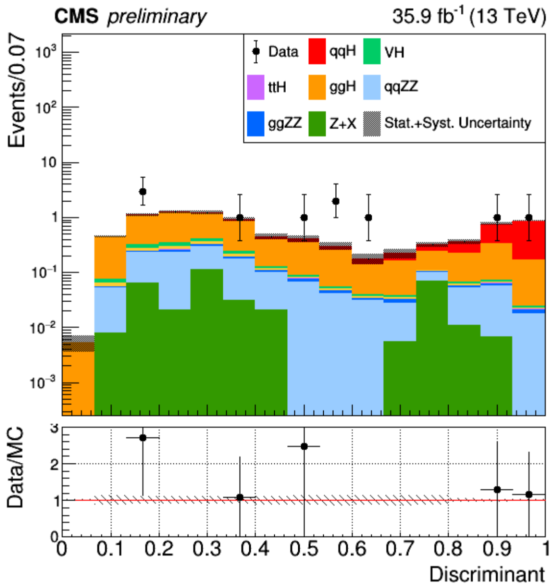
Table 8.3: The expected event yields from background and signal are shown with the number of observed events after the VBF-SR selections for the *Njets2* category of the Run 2 dataset. The quoted uncertainties include statistical and systematic ones.

Process	4μ	4e	$2e2\mu$	4l
ggH	3.21±0.12	1.53±0.13	3.90±0.23	8.65±0.30
VH	0.45±0.01	0.23±0.01	0.56±0.02	1.24±0.03
ttH	0.07±0.00	0.05±0.01	0.07±0.01	0.20±0.01
qqZZ+ZZJJ	0.48±0.02	0.18±0.02	0.50±0.03	1.16±0.04
ggZZ	0.11±0.01	0.05±0.01	0.04±0.00	0.20±0.01
Z+X	0.42±0.07	0.04±0.01	0.24±0.03	0.70±0.08
$\sum$ backgrounds	4.74±0.14	2.09±0.14	5.32±0.24	12.15±0.31
qqH (signal $m_H = 125GeV$)	0.72±0.52	0.37±0.37	0.92±0.72	2.01±0.96
Total expected	5.46±0.54	2.46±0.39	6.24±0.76	14.16±1.01
Observed	2	2	4	8

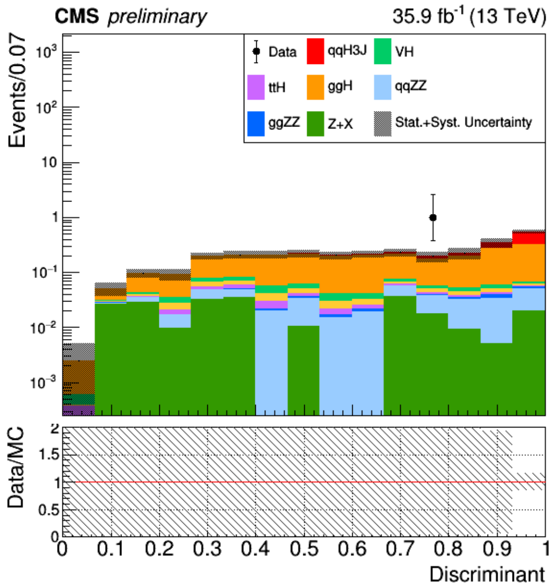
Table 8.4: The expected event yields from background and signal are shown with the number of observed events after the VBF-SR selections for the *Njets3* category of the Run 2 dataset. The quoted uncertainties include statistical and systematic ones.

Process	4μ	4e	$2e2\mu$	4l
ggH	11.30±0.34	5.32±0.35	13.76±0.62	30.38±0.79
VH	1.20±0.02	0.61±0.03	1.46±0.04	3.27±0.05
ttH	0.10±0.00	0.07±0.01	0.09±0.01	0.26±0.01
qqZZ+ZZJJ	2.33±0.06	1.02±0.07	2.64±0.12	5.99±0.16
ggZZ	0.42±0.03	0.20±0.02	0.23±0.02	0.86±0.04
Z+X	0.72±0.09	0.21±0.04	0.71±0.07	1.63±0.12
$\sum$ backgrounds	16.06±0.36	7.43±0.36	18.89±0.64	42.38±0.82
qqH (signal $m_H = 125GeV$)	3.29±0.46	1.65±0.37	4.07±0.70	9.01±0.91
Total expected	19.35±0.59	9.08±0.52	22.96±0.95	51.39±1.23
Observed	20	10	19	49

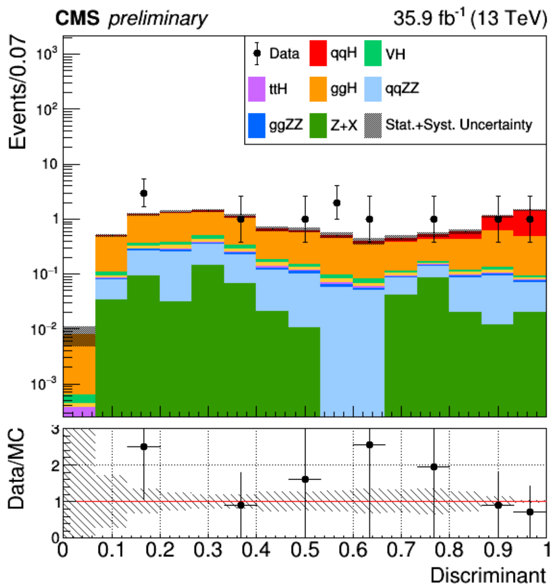
Table 8.5: The expected event yields from background and signal are shown with the number of observed events after the VBF-SR selections for the *Njets2* + *Njets3* category of the Run 2 dataset. The quoted uncertainties include statistical and systematic ones.



(a)

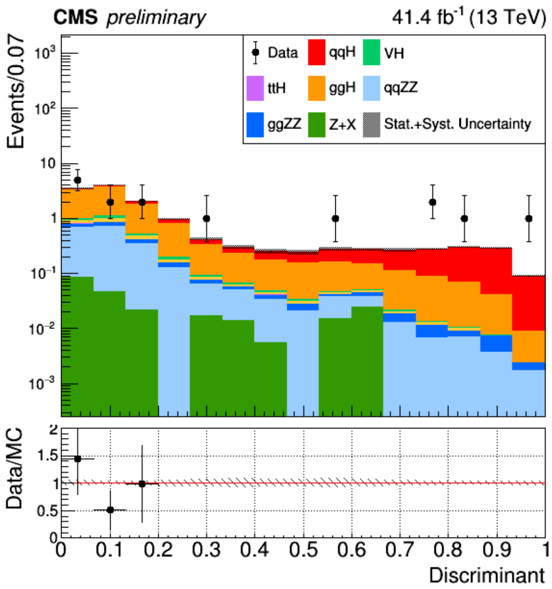


(b)

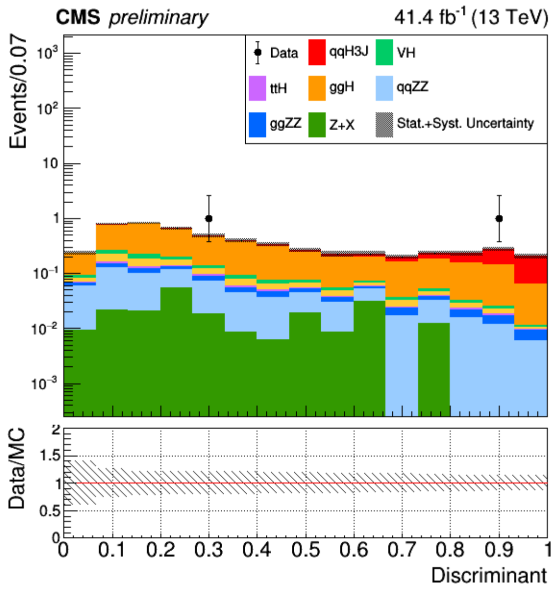


(c)

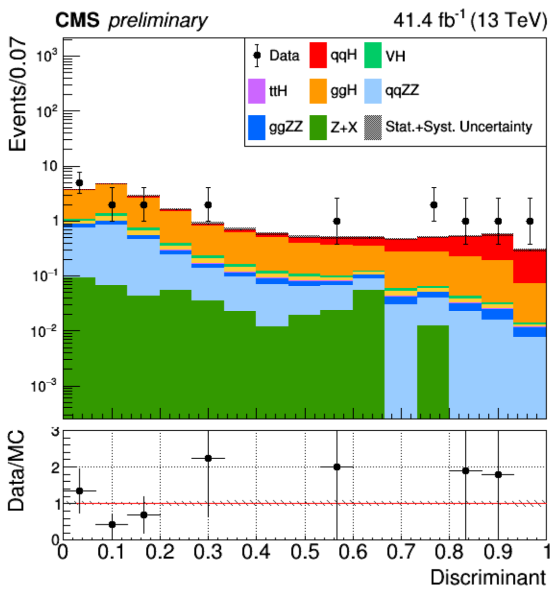
Figure 8.5: Pre-fit distributions of the neural network developed for *Njets2* (a), *Njets3* (b) categories and the combination of the two categories (c). The plots shown are for the 2016 dataset.



(a)

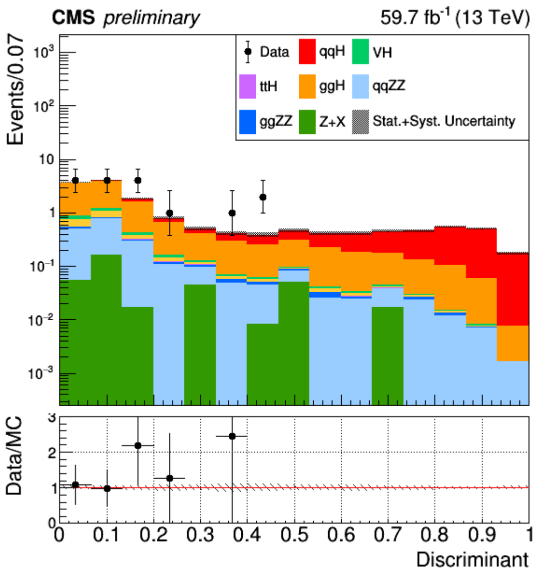


(b)

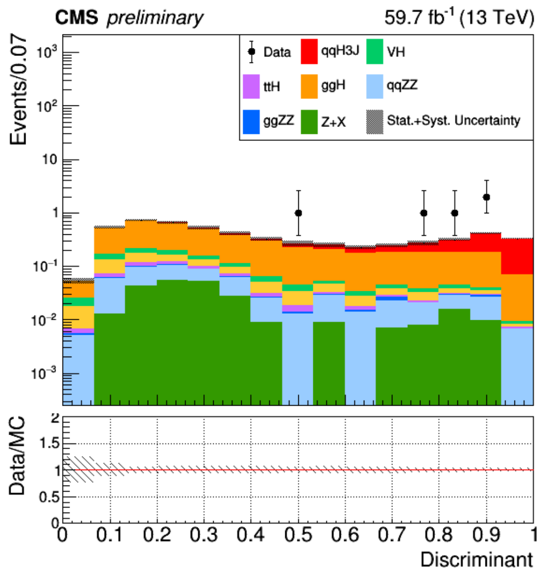


(c)

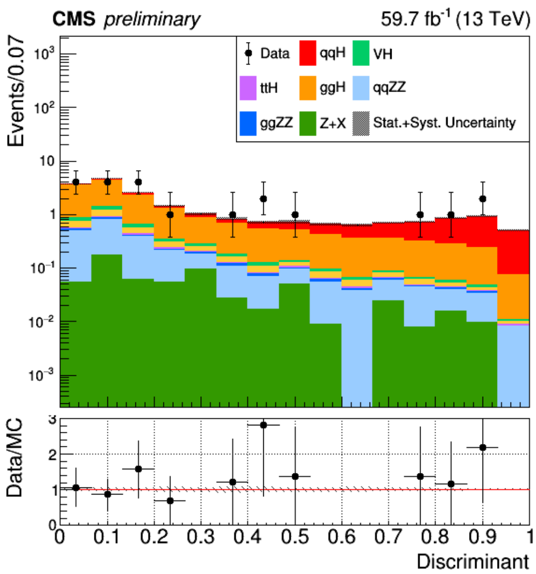
Figure 8.6: Pre-fit distributions of the neural network developed for *Njets2* (a), *Njets3* (b) categories and the combination of the two categories (c). The plots shown are for the 2017 dataset.



(a)

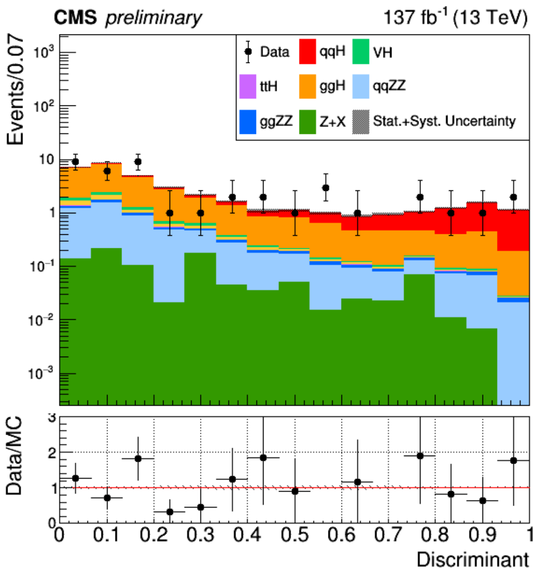


(b)

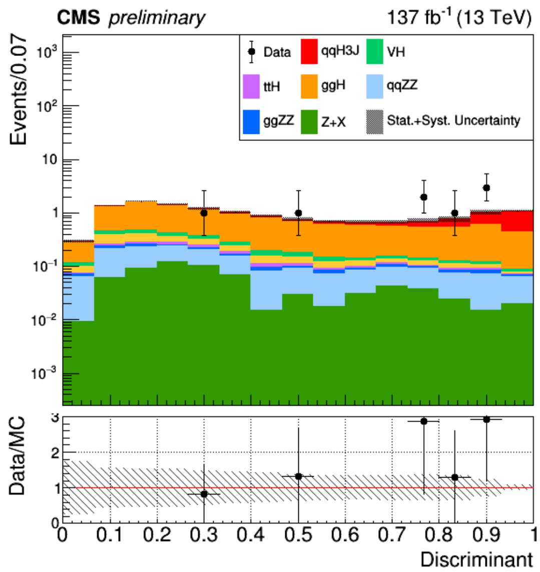


(c)

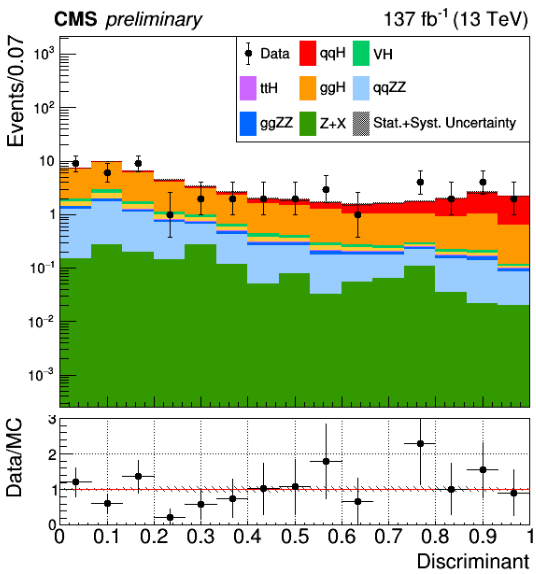
Figure 8.7: Pre-fit distributions of the neural network developed for *Njets2* (a), *Njets3* (b) categories and the combination of the two categories (c). The plots shown are for the 2018 dataset.



(a)



(b)



(c)

Figure 8.8: Pre-fit distributions of the neural network developed for *Njets2* (a), *Njets3* (b) categories and the combination of the two categories (c). The plots shown are for the Run 2 dataset.

8.3.2 Validation of the statistical model

The validation of the statistical model and the nuisance parameters is done by testing the impact of the nuisance parameters on the measurement of the signal strengths and by evaluating the χ^2 of the global fit. As no anomalous behaviours are seen, the observed data events are compared to the yields predicted by the signal and the background simulated processes in the SM scenario for each category (*Njet2* and *Njet3*) and their combination. A global fit performed to all these distributions yields the measurement of the signal strength of the qqH signal process and the subsequent estimation of the statistical and systematic uncertainties on this value.

8.3.2.1 Impact of systematic uncertainties

The sources of systematic uncertainties, described in Section 8.1, yield independent nuisance parameters that can affect the shape and normalization of the distributions of the discriminant observables. Their impact on measuring the qqH signal strength was assessed by evaluating the so-called impacts and pulls. The impact of a nuisance parameter θ is defined as the shift $\Delta\mu$ induced on the value of signal strength by fixing θ to its post-fit value $\hat{\theta} \pm 1\sigma$, while the other nuisance parameters are estimated as the minimum of a profiled likelihood; the procedures address how the estimated signal strength changes due to the floating systematic uncertainty. The pulls are defined as the difference between the best fit estimation $\hat{\theta}$ from the likelihood maximization and the initial value of the nuisance parameter, normalized to the value of the uncertainty. The impacts and pulls of the leading 30 systematic uncertainties present in this analysis, combining the *Njets2* and *Njets3* categories for the three years of data taking are shown in Fig 8.9 - Fig 8.11 and for the combined Run 2 datasets are shown in Fig 8.12. No parameter is significantly pulled away from its nominal value or overconstrained by the fit procedure, indicating the quality of the statistical model. The most important source of systematic uncertainty that affects the measurements for 2016 and 2017 is the systematics on the 3^{rd} jet, while for 2018 and the combined Run 2 is the systematics on the jet energy scale for the leading jet.

8.3.2.2 Goodness-of-fit test

The goodness-of-fit test further verifies the correct data modelling with the saturated method [166]. It is defined as a generalization of the χ^2 method for data that are not normally distributed, as in the case of the number of events in the binned distributions of the analysis. The test is built as a likelihood ratio, but the alternative hypothesis is considered to be the one that exactly matches the data. The distribution of this test statistic is obtained from pseudo-experiments generated from the modelled processes and compared to the distribution calculated using the observed data. The comparison is shown in Fig. 8.13, where the observed value of the χ^2 goodness of fit, with a p-value within the distribution of the 5000 generated pseudo-experiments, concludes that the statistical model describes well the observations.

8.3.3 Post-fit yields and distributions

The expected and observed event yields for the signals and backgrounds in the signal regions of the analysis are reported in Tables. 8.6 - 8.8 for each jet category (*Njets2*, *Njets3*) and for the combination, respectively. The signal and background yields in the table are obtained after the

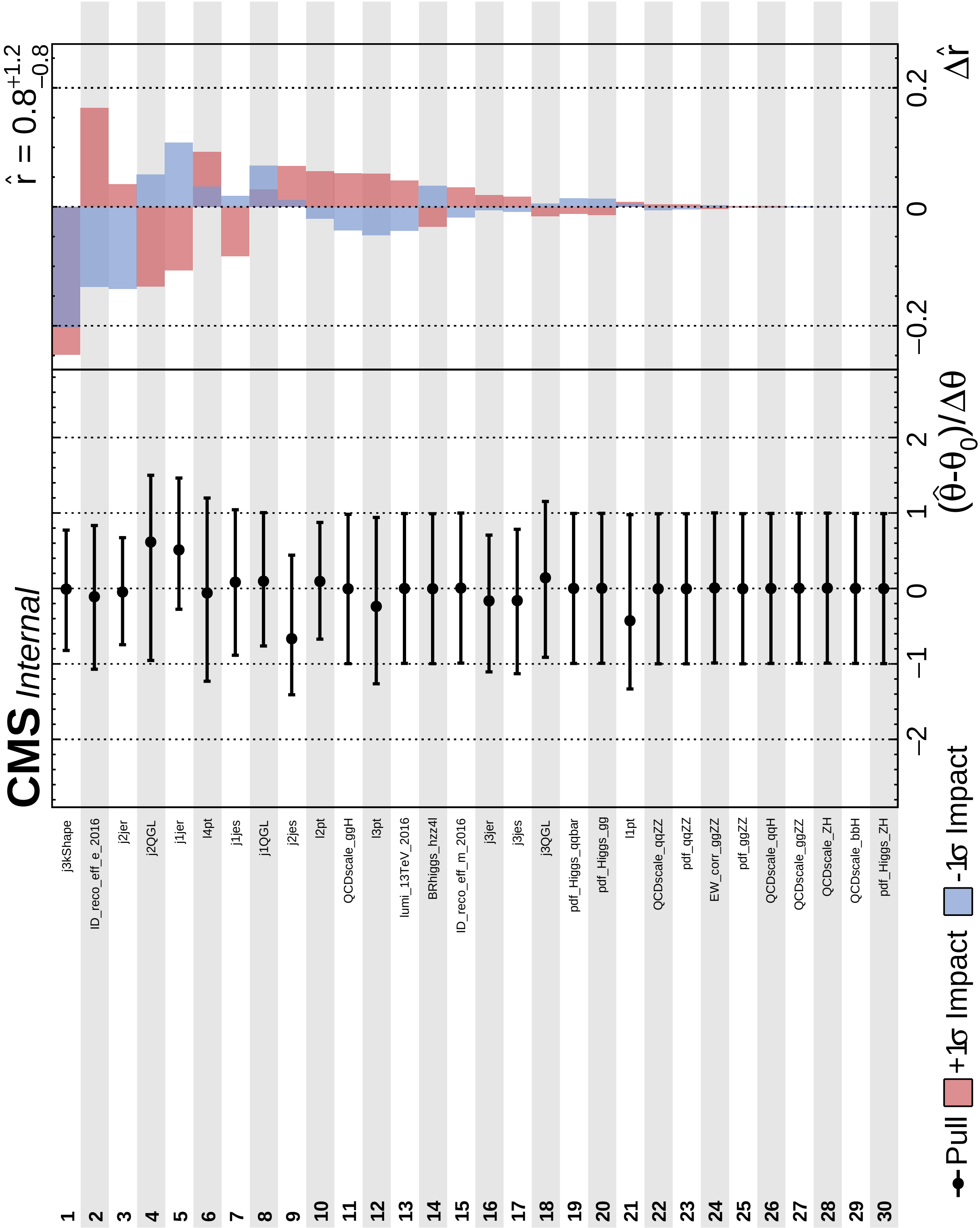


Figure 8.9: List of the 30 nuisances parameters with the highest impact on the measurement of the qqH signal strength. The impact plot was obtained with the combination of the two categories. The left row shows the pull of the nuisance parameters, while the right one shows the impact of the nuisances on the measurement of the signal strength (μ_{qqH}) for the 2016 year.

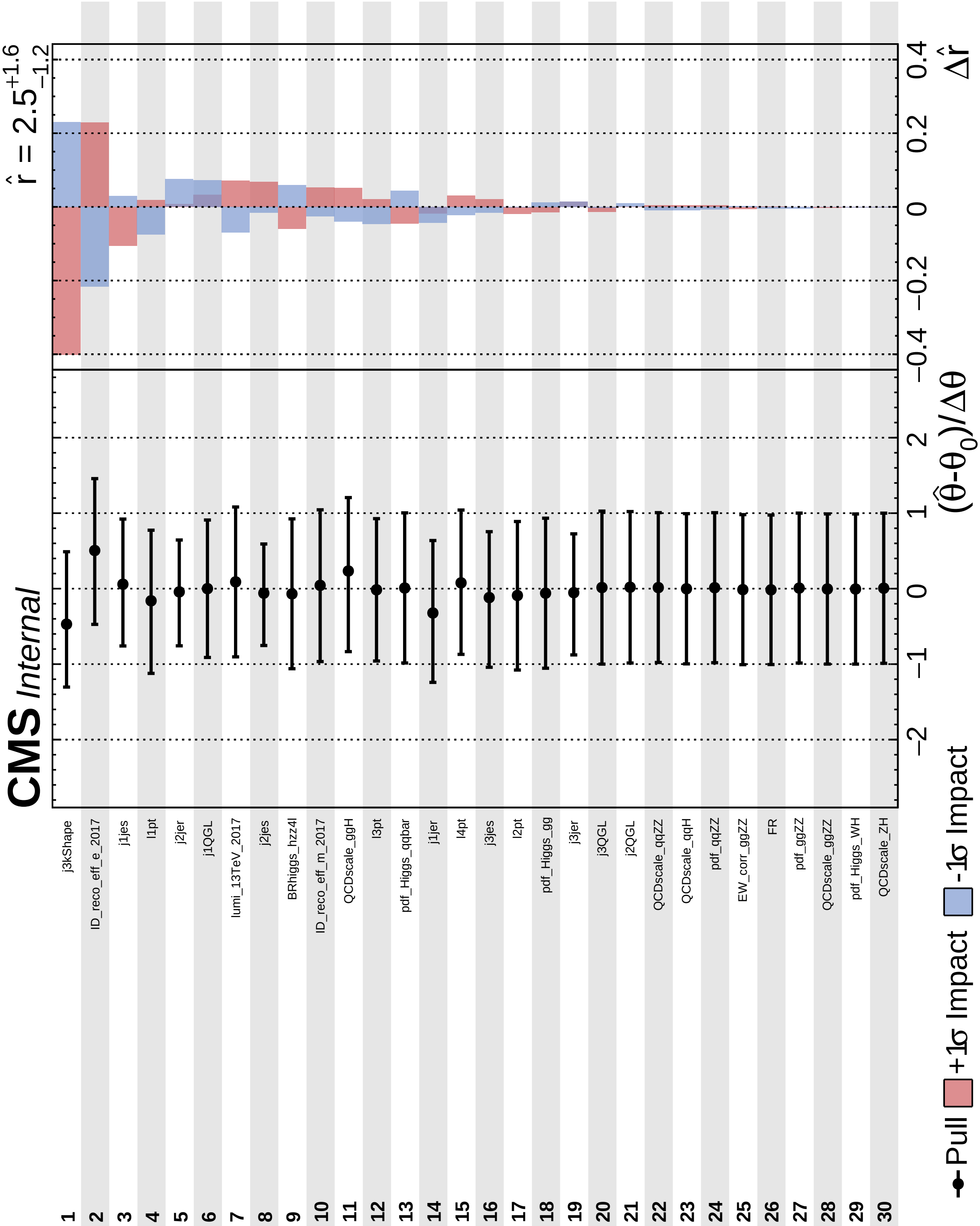


Figure 8.10: List of the 30 nuisances parameters with the highest impact on the measurement of the qqH signal strength. The impact plot was obtained with the combination of the two categories. The left row shows the pull of the nuisance parameters, while the right one shows the impact of the nuisances on the measurement of the signal strength (μ_{qqH}) for the 2017 year.

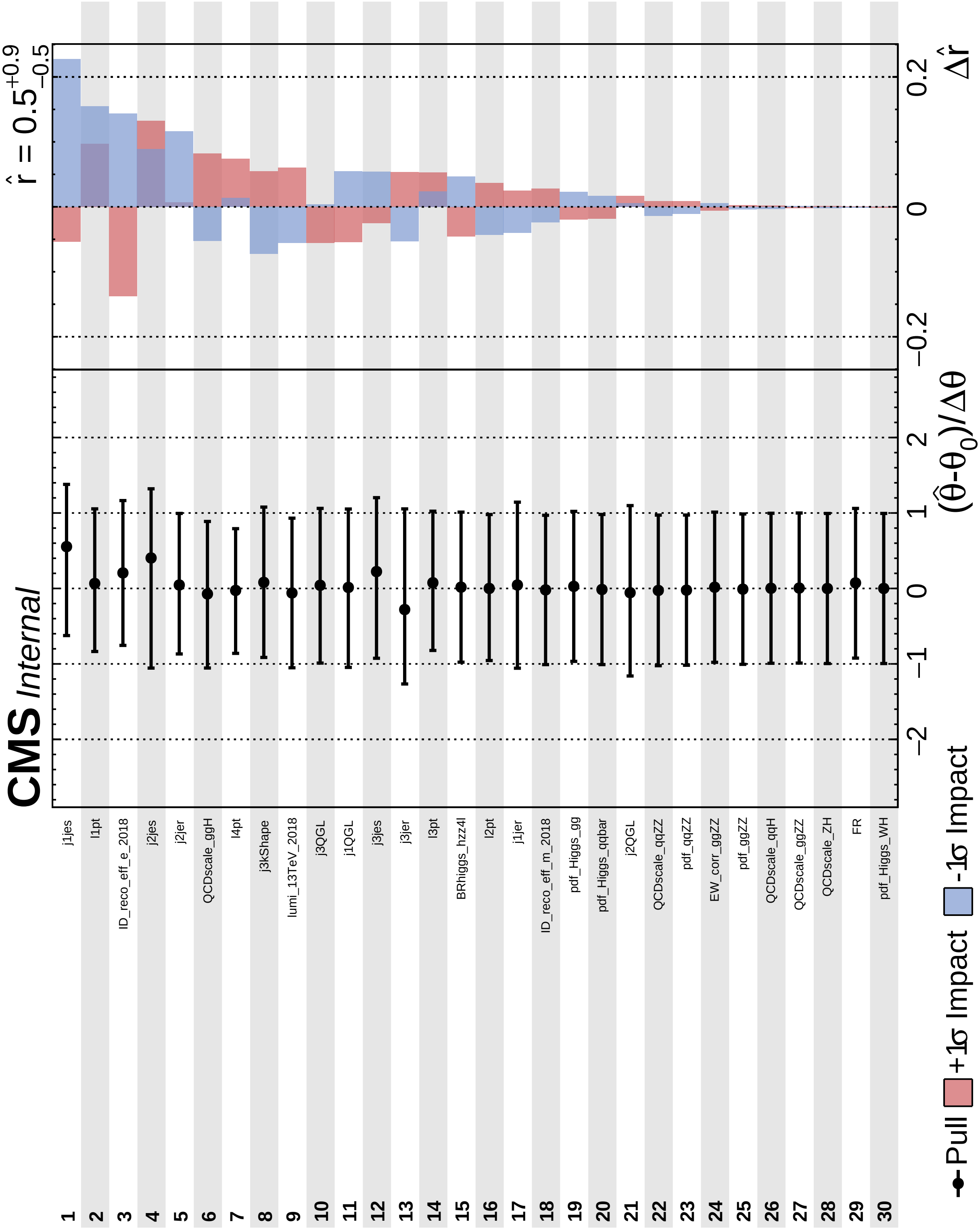


Figure 8.11: List of the 30 nuisances parameters with the highest impact on the measurement of the qqH signal strength. The impact plot was obtained with the combination of the two categories. The left row shows the pull of the nuisance parameters, while the right one shows the impact of the nuisances on the measurement of the signal strength (μ_{qqH}) for the 2018 year.

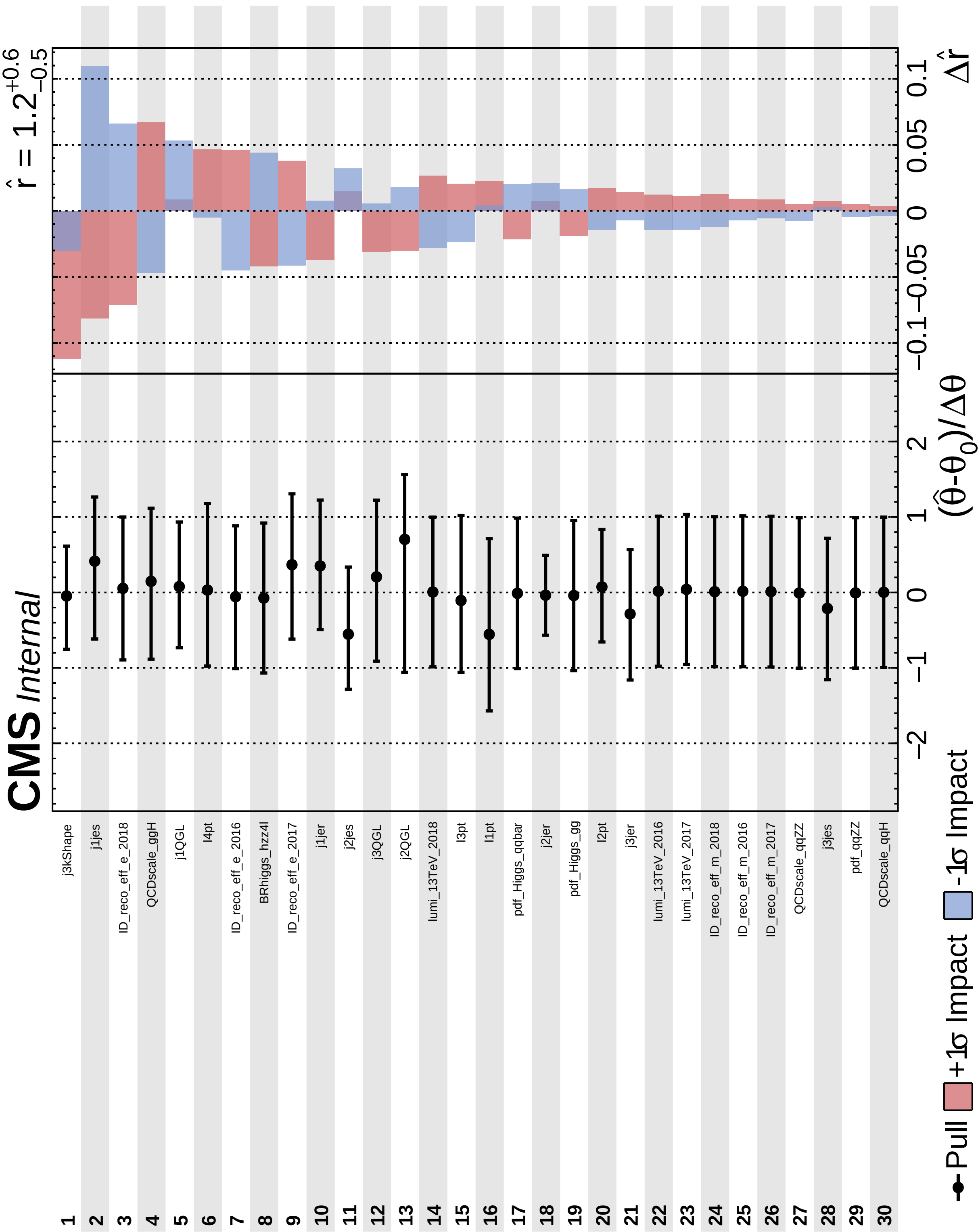


Figure 8.12: List of the 30 nuisances parameters with the highest impact on the measurement of the qqH signal strength. The impact plot was obtained with the combination of the two categories. The left row shows the pull of the nuisance parameters, while the right one shows the impact of the nuisances on the measurement of the signal strength (μ_{qqH}) for Run II.

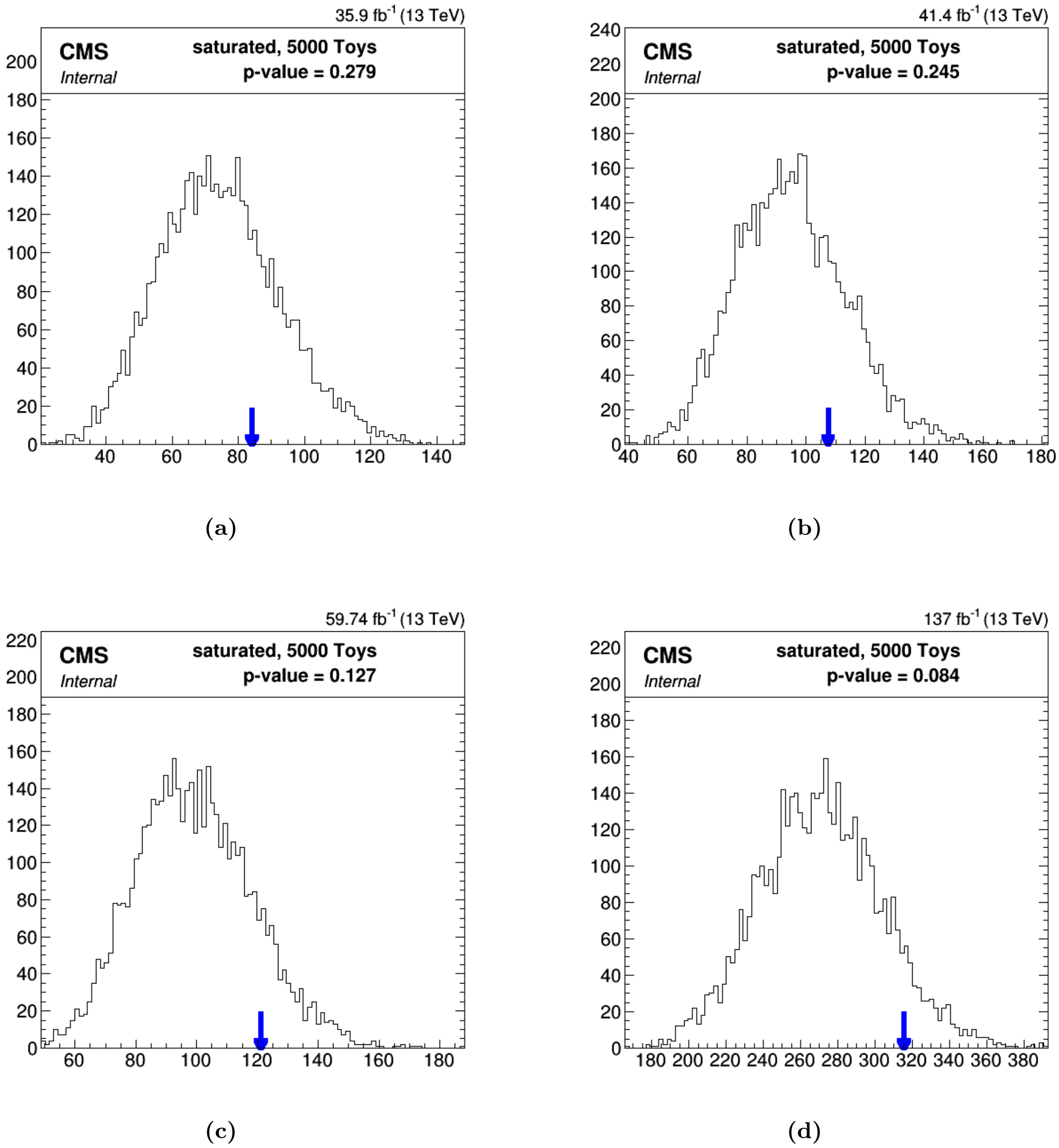


Figure 8.13: Distribution of the goodness of fit variable χ^2 estimated with the saturated model on 5000 pseudo-experiments "Toys" (black histogram) compared to the observed value (blue arrow) for 2016 (a), 2017 (b), 2018 (c), and for the combined Run 2 (d).

maximum likelihood fit, i.e. post-fit yields, which means that they are computed for the best fit values of nuisance parameters. The results show that the number of data events and the expected signal and background yields are in good agreement.

The post-fit distributions of the neural networks discriminant in each of the *Njets2* and *Njets3* categories for each year and the combined Run2 are shown in Fig(s). 8.14 - 8.17(a), (b), while for their combination are shown in Figs. 8.14 - 8.17 (c). The plots confirm a good agreement in shape between the distributions of observed and expected events.

Process	4μ	4e	$2e2\mu$	4l
ggH	8.14±0.30	3.80±0.35	9.89±0.61	21.84±0.76
VH	0.75±0.01	0.38±0.03	0.90±0.04	2.03±0.05
ttH	0.02±0.00	0.02±0.00	0.02±0.00	0.05±0.00
qqZZ+ZZJJ	1.85±0.06	0.83±0.07	2.15±0.13	4.83±0.16
ggZZ	0.32±0.02	0.15±0.02	0.19±0.02	0.66±0.04
Z+X	0.30±0.05	0.17±0.04	0.47±0.06	0.95±0.09
$\sum$ backgrounds	11.38±0.31	5.35±0.36	13.62±0.63	30.34±0.79
qqH (signal $m_H = 125GeV$)	2.93±0.91	1.45±0.45	3.59±1.10	7.97±1.50
Total expected	14.31±0.96	6.80±0.58	17.20±1.27	38.31±1.69
Observed	18	8	15	41

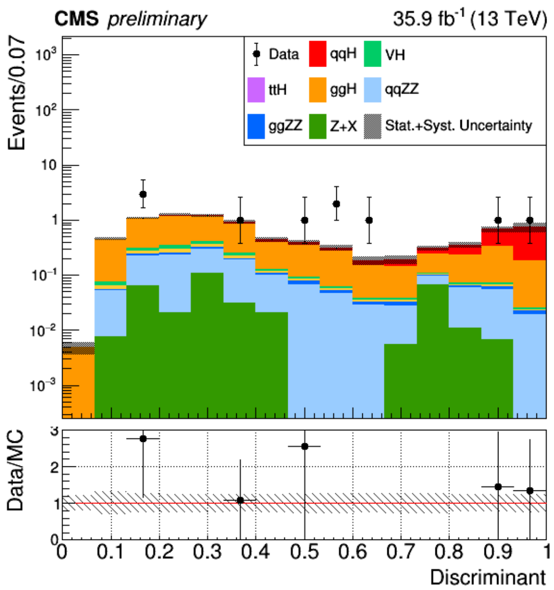
Table 8.6: Background and signal estimations, derived from the maximum likelihood fit for the s+b model to the observed data for the Njet2 category of Run 2 dataset. The quoted uncertainties include the statistical and systematic components.

Process	4μ	4e	$2e2\mu$	4l
ggH	3.13±0.12	1.47±0.13	3.77±0.23	8.38±0.29
VH	0.45±0.01	0.23±0.01	0.55±0.02	1.23±0.03
ttH	0.07±0.00	0.05±0.01	0.07±0.01	0.20±0.01
qqZZ+ZZJJ	0.48±0.02	0.18±0.02	0.49±0.03	1.15±0.04
ggZZ	0.11±0.01	0.05±0.01	0.04±0.00	0.20±0.01
Z+X	0.40±0.07	0.03±0.01	0.23±0.03	0.66±0.08
$\sum$ backgrounds	4.64±0.14	2.02±0.13	5.16±0.23	11.81±0.30
qqH (signal $m_H = 125GeV$)	0.98±0.57	0.50±0.31	1.25±0.75	2.73±0.99
Total expected	5.61±0.59	2.52±0.34	6.41±0.79	14.54±1.04
Observed	2	2	4	8

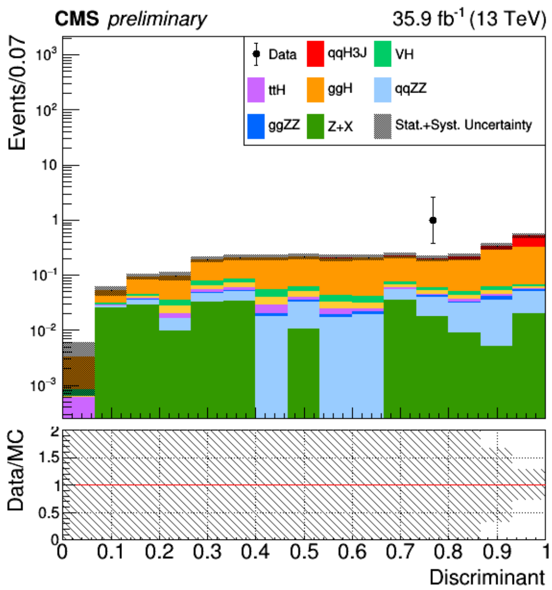
Table 8.7: Background and signal estimations, derived from the maximum likelihood fit for the s+b model to the observed data for the Njet3 category of Run 2 dataset. The quoted uncertainties include the statistical and systematic components.

8.3.4 Measurement of the signal strength

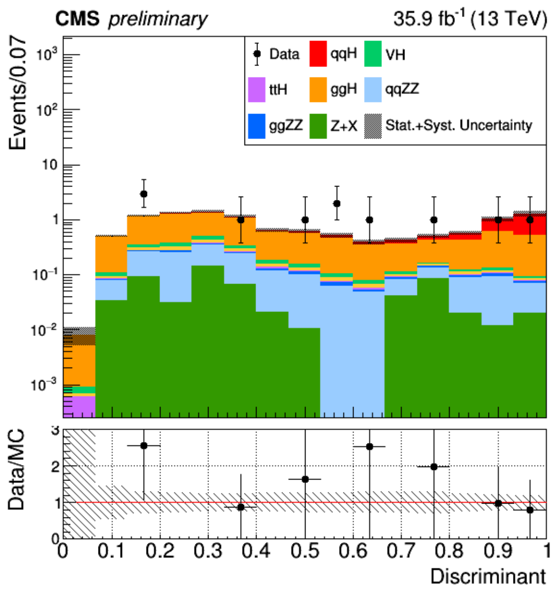
A simultaneous fit is performed to extract the signal strength modifier for the VBF Higgs signal, defined as the ratio of the observed VBF Higgs signal yield to the expectation from the standard model. The likelihood ratio (see eq. 8.5) is performed with $\mu = 1$, that describe the SM expected VBF signal, and $\hat{\mu}$ is left to float during the fit on the observed events, that describe the VBF signal counted from observed data. In this analysis the Higgs VBF signal strength has been measured to



(a)



(b)



(c)

Figure 8.14: Post-fit distributions of the neural network developed for *Njets2* (a), *Njets3* (b) categories and the combination of the two categories (c). The fit is done with the assumption of the S+B hypothesis. The plots shown are for the 2016 dataset.

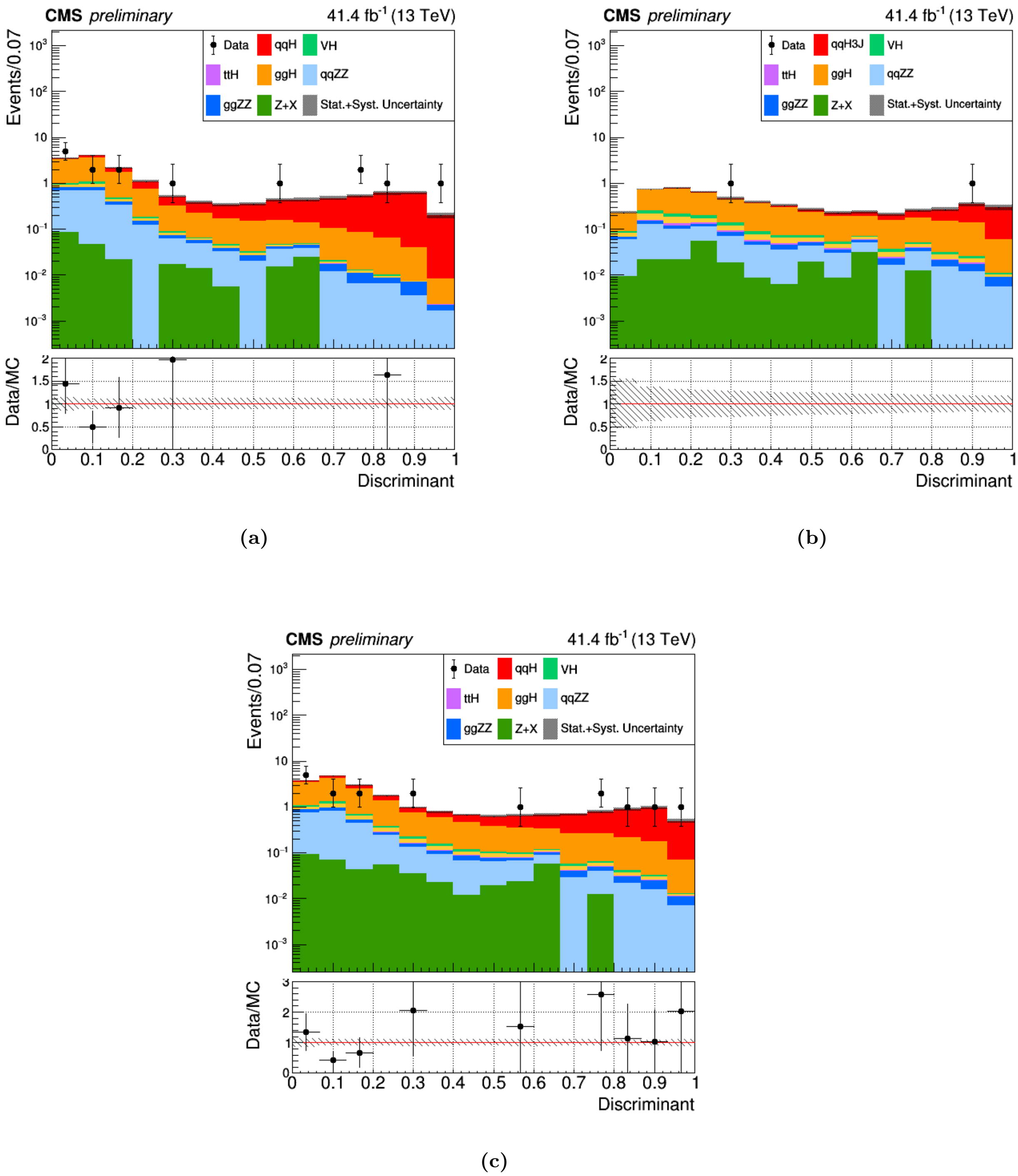
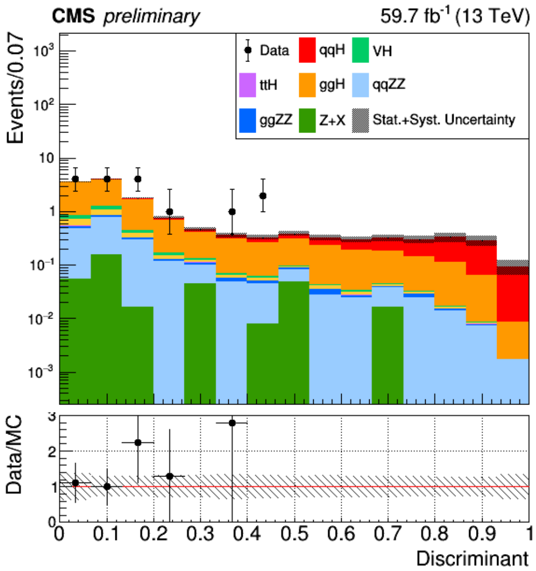
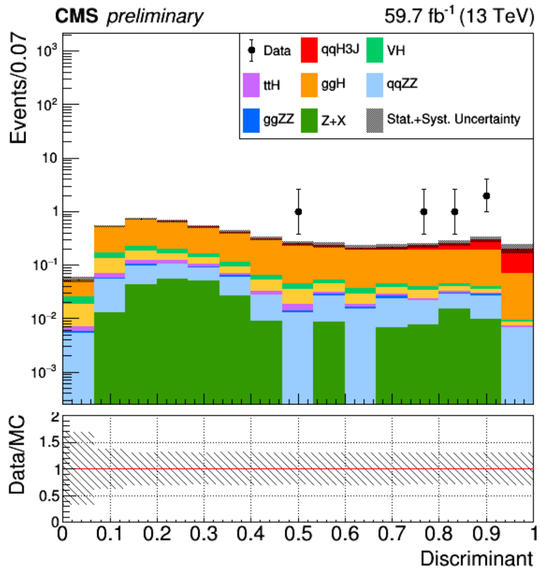


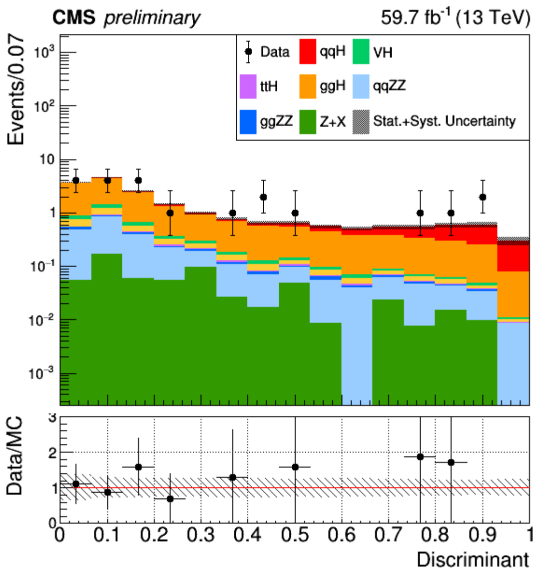
Figure 8.15: Post-fit distributions of the neural network developed for *Njets2* (a), *Njets3* (b) categories and the combination of the two categories (c). The fit is done with the assumption of the S+B hypothesis. The plots shown are for the 2017 dataset.



(a)

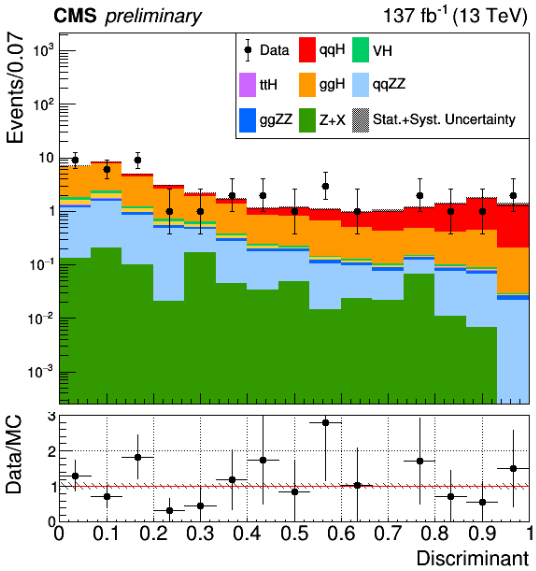


(b)

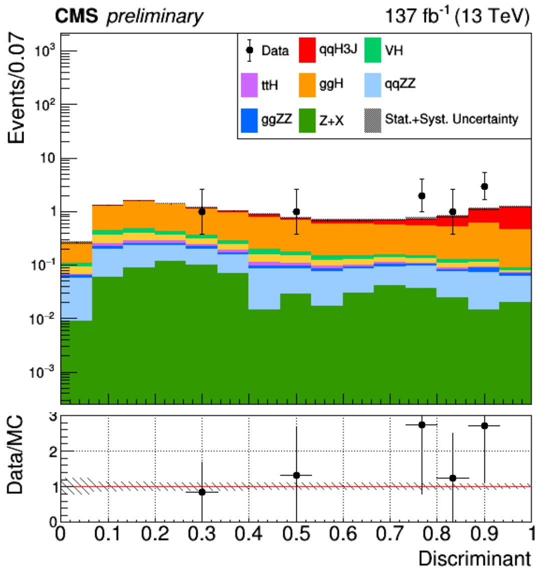


(c)

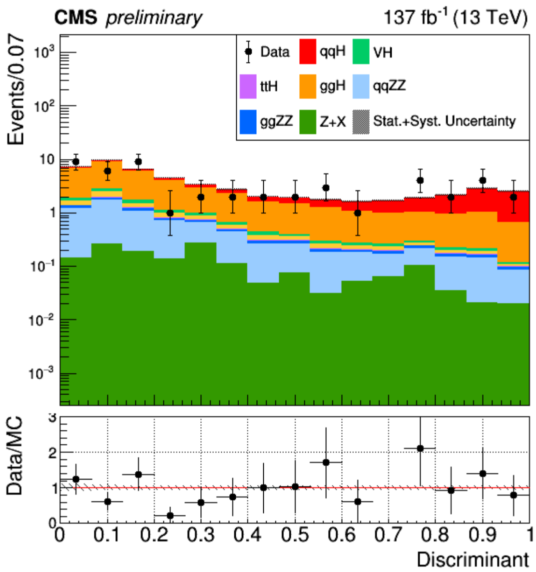
Figure 8.16: Post-fit distributions of the neural network developed for *Njets2* (a), *Njets3* (b) categories and the combination of the two categories (c). The fit is done with the assumption of the S+B hypothesis. The plots shown are for the 2018 dataset.



(a)



(b)



(c)

Figure 8.17: Post-fit distributions of the neural network developed for *Njets2* (a), *Njets3* (b) categories and the combination of the two categories (c). The fit is done with the assumption of the S+B hypothesis. The plots shown are for the Run 2 dataset.

Process	4μ	4e	$2e2\mu$	4l
ggH	11.18±0.34	5.19±0.34	13.48±0.61	29.85±0.77
VH	1.19±0.02	0.60±0.03	1.44±0.04	3.24±0.05
ttH	0.10±0.00	0.07±0.01	0.09±0.01	0.25±0.01
qqZZ+ZZJJ	2.32±0.06	1.00±0.07	2.61±0.12	5.94±0.15
ggZZ	0.42±0.03	0.20±0.02	0.23±0.02	0.85±0.04
Z+X	0.69±0.09	0.20±0.05	0.69±0.07	1.58±0.12
$\sum$ backgrounds	15.91±0.35	7.26±0.35	18.54±0.62	41.71±0.80
qqH (signal $m_H = 125GeV$)	3.92±0.92	1.95±0.47	4.83±1.13	10.70±1.53
Total expected	19.83±0.99	9.21±0.58	23.37±1.29	52.41±1.73
Observed	20	10	19	49

Table 8.8: Background and signal estimations, derived from the maximum likelihood fit for the s+b model to the observed data for Njet2 + Njet3 category of Run 2 dataset. The quoted uncertainties include the statistical and systematic components.

be $\mu_{qqH} = \sigma^{Obs}_{qqH}/\sigma^{SM}_{qqH} = 1.19^{+0.62}_{-0.50}$, for the SM Higgs boson with mass 125 GeV and the collected Run 2 data which amount to 137 fb^{-1}. The uncertainties on the measurement of the signal strength are comparable with the expected values $\mu^{exp}_{qqH} = 1.00^{+0.54}_{-0.43}$. The results are shown in Fig. 8.18 for the observed and expected profile likelihood scans of the two jet based categories used in the analysis and for their combination (a) and the summary of the minimum likelihood ratio values and the 68% CL for each category, year, and their combination (b). The corresponding numerical values are summarized in Table 8.9.

	$\mu^{2016,2J}_{qqH}$	$\mu^{2016,3J}_{qqH}$	$\mu^{2017,2J}_{qqH}$	$\mu^{2017,3J}_{qqH}$	$\mu^{2018,2J}_{qqH}$	$\mu^{2018,3J}_{qqH}$	$\mu^{Run2,combined}_{qqH}$
Expected	$1.00^{+1.29}_{-0.79}$	$1.00^{+4.01}_{-1.00}$	$1.00^{+1.31}_{-0.81}$	$1.00^{+4.55}_{-1.00}$	$1.00^{+0.91}_{-0.63}$	$1.00^{+1.92}_{-0.97}$	$1.00^{+0.54}_{-0.43}$
Observed	$1.12^{+1.50}_{-0.97}$	$0.0^{+1.61}_{-1.38}$	$2.78^{+1.67}_{-1.26}$	$0.0^{+2.74}_{-0.0}$	$0.0\ ^{+0.39}_{-0.00}$	$2.93^{+2.50}_{-1.77}$	$1.19^{+0.62}_{-0.50}$

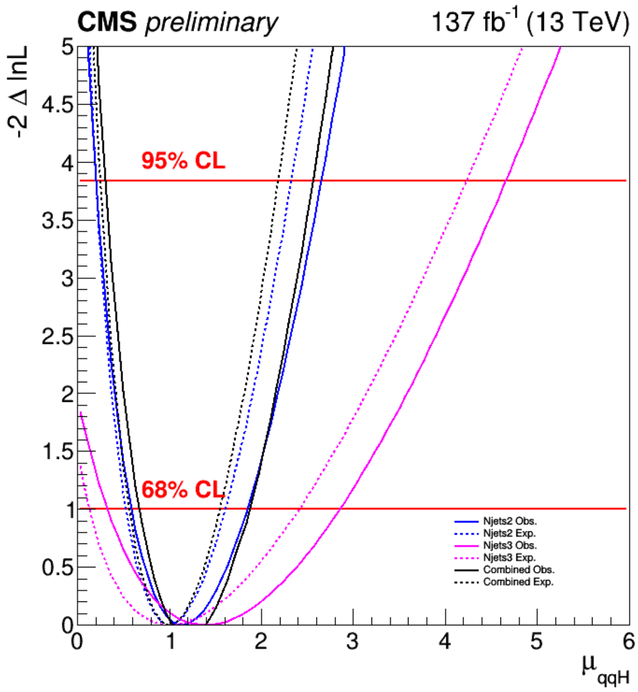
Table 8.9: Expected and observed signal strength modifiers from each category of the number of jets and the combined Run 2 with 137 fb^{-1} at $\sqrt{s} = 13$ TeV.

8.3.5 VBF Higgs production significance

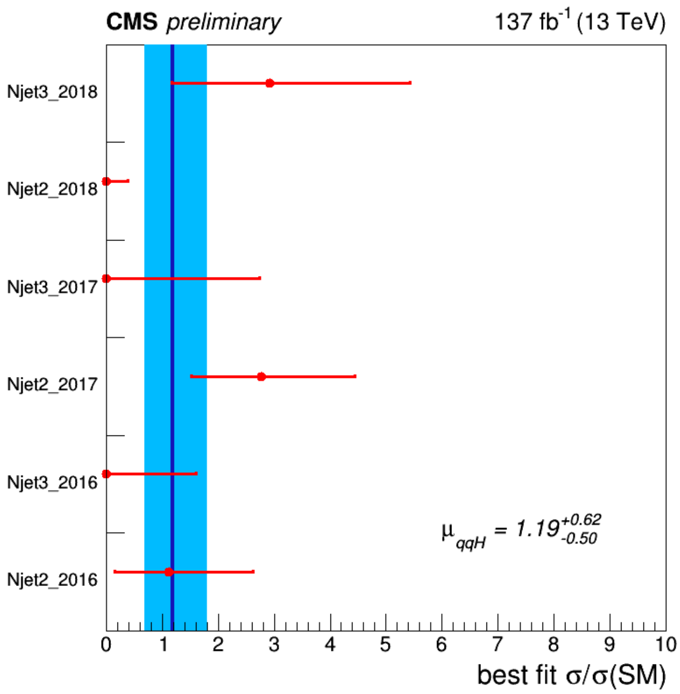
In order to quantify the observation of the VBF Higgs signal in each category and their combination, their significance has been computed, as explained in Section 8.2.3. Fig. 8.19 shows the expected and observed significances for each jet-based category and the combination. The observed (expected) significance amounts to $\sigma^{obs}_{qqH} = 3.02\sigma$ ($\sigma^{exp}_{qqH} = 3.06\sigma$), equivalent to P-value $p^{obs}_0 = 1.2 \times 10^{-3}$ ($p^{exp}_0 = 1.1 \times 10^{-3}$). The achieved sensitivity confirms the evidence of the VBF Higgs production (see Section 8.2.3).

8.3.6 Projection for Future Luminosity Scenarios

According to the LHC operation schedule shown in Fig 2.2, the LHC is foreseen to operate in 2022 after the long shutdown "LS2"; this operation period is known as "Run III", with an expected integrated luminosity of 350 fb^{-1}. This data taking period will be followed by a long shutdown for the LHC upgrade, referred to as "LS3". After LS3, LHC operation will resume to increase the



(a)



(b)

Figure 8.18: The best fit value found, given the observed data, for μ_{qqH}. In (a) is shown the observed and expected profile likelihood scans of the qqH signal strength modifier from each category of the number of jets and the combined Run 2. In (b) is shown the observed signal strength for each category of the number of jets, for each year in run 2, and for the combination with the associated uncertainty.

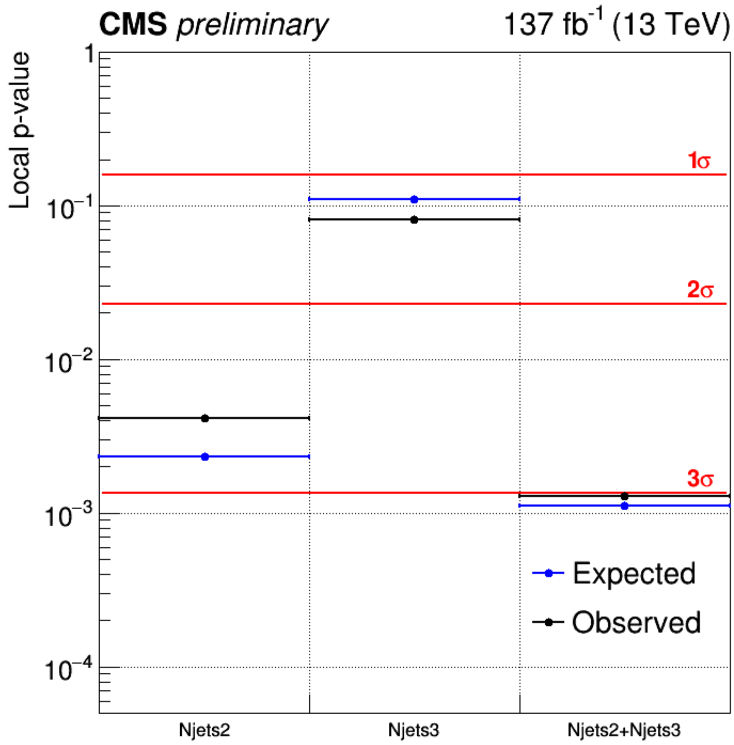


Figure 8.19: The significance of the local excess w.r.t the SM background expectation for each jet-based category and their combination.

integrated luminosity by a factor of ten to enter the so-called High Luminosity Phase of LHC HL-LHC. This section presents a simple projection of the expected significance evolution of the present analysis with the future integrated luminosities planned to be achieved at the LHC Run 3 and HL-LHC. This study does not consider the increase of the production cross-section of the physics processes after increasing the center of mass-energy from 13 to 14 TeV. Also, no approach used to scale the uncertainties affect the analysis was applied since that demands a more dedicated study of the uncertainties dependencies beyond this analysis's scope. According to this projection, it is expected to have enough significance for the VBF Higgs production observation with the luminosity planned for the CMS Run 3 (350 fb^{-1}).

Dataset	2016	2017	2018	Run 2	Run 3	HL-LHC
Luminosity (fb^{-1})	35.9	41.4	59.74	137	350	3000
Expected significance	1.52	1.60	2.30	3.1	5.03	14.74

Table 8.10: The expected VBF Higgs significance from the present luminosity and the expected future scenarios at the LHC.

Conclusions

The analysis described in this thesis presents a dedicated measurement of the Higgs boson production in the VBF production mode in the $H \rightarrow ZZ \rightarrow 4l(l = e,\ \mu)$ channel. The dataset used in the analysis corresponds to proton-proton collision data produced at the LHC at a center-of-mass energy of 13 TeV and collected by the CMS detector equivalent to an integrated luminosity of $137 fb^{-1}$.

Several developments were done to the different aspects of the analysis, and new techniques were used to improve on the previous measurement published by CMS obtained with Run 2 data (see Section 1.5). Deep neural networks are used as a discriminant for the VBF Higgs signal. In addition, further gains come from the effort devoted to optimizing the multivariate discriminator to separate the signal from the background as best as possible. The analysis proposes the usage of the 3^{rd} jet when available in the event; therefore, a new sample for VBF Higgs production associated with three hard jets at NLO was generated to be used for the first time in the VBF Higgs analysis. Two jet-based categories have been defined in order to optimize the analysis and adequately handle the usage of the third jet when it is available in the events. For each category, a NN discriminant was developed.

The measured cross-section is consistent with the SM expectation, with a signal strength $\mu_{qqH} = \sigma^{Obs}_{qqH}/\sigma^{SM}_{qqH} = 1.19^{+0.62}_{-0.50}$. The excess of events observed over the background expectation corresponds to a significance of $\sigma^{obs}_{qqH} = 3.02\sigma$; thus, it constitutes the evidence of the Higgs boson production via vector boson fusion. The projection for future luminosity is expected to have enough significance for the VBF Higgs production observation (5σ) with the expected integrated luminosity of LHC Run 3 (350 fb^{-1}).

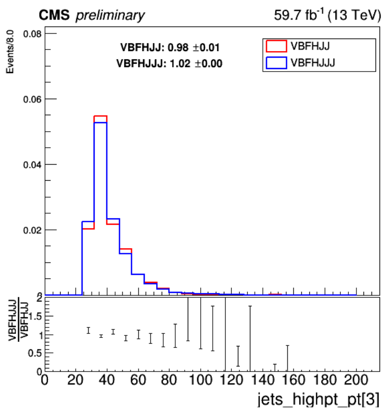
The muon tracking performance is crucial for the overall performance of the CMS experiment, and it affects the reconstruction performance for all the objects used in the VBF Higgs analysis, which is the main topic for this thesis. Therefore, a study was conducted to measure the performance of the CMS muon tracking with the tag-and-probe technique exploiting the $Z \rightarrow \mu^+\mu^-$ resonance using CMS Run 2 data with an integrated luminosity of $137 fb^{-1}$. As a result, a tracking efficiency of about 99.9 % was guaranteed for all reconstructed muon trajectories (all-tracks) in the entire muon pseudorapidity acceptance through the whole Run 2 data taking. Also a tracking efficiency for the subset of trajectories in which the CMS Tracker only is used to seed the measurement (tracker-only seeded tracks) of about 98% is guaranteed.

One of the detectors that play a critical role in triggering muons in the CMS experiment is the RPC detector. Hence, a study was performed to evaluate the effects on the RPC efficiency, cluster size and detector intrinsic noise rate of variations of the electronics threshold voltage. The results obtained are in agreement with the expectations: a decrease in the applied discrimination thresholds increases the RPC efficiency, which is a benefit for the system but, at the same time, induces an increase in cluster size and intrinsic noise, which could deteriorate spatial resolution and accidentals, respectively. Nevertheless, a $\approx 0.9\%$ gain in efficiency consequent to a 5 mV decrease of the threshold with respect to the default one is to be preferred w.r.t. the corresponding minimal increase in cluster size, of around 0.07 strips, and intrinsic noise, roughly 0.04 Hz/cm^2. This would induce to consider thoughtfully to adopt such a solution when the CMS RPC system is restarted after the current shutdown.

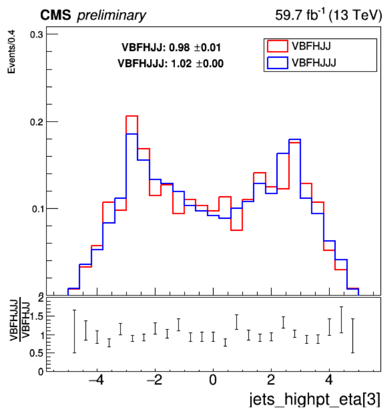
Appendix A

The 3^{rd} jet in the two signal samples VBF-H2J and VBF-H3J

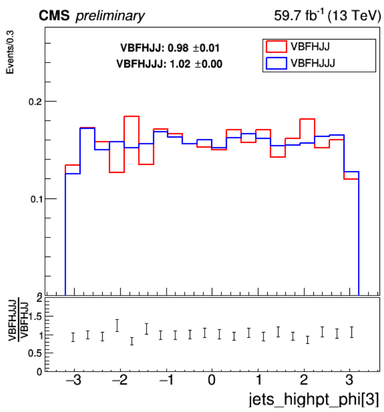
The appendix contains comparison plots between the 3^{rd} jet from the two signal samples VBF-H2J, which has two jets at NLO and VBF-H3J, which has three jets at NLO for different observables.



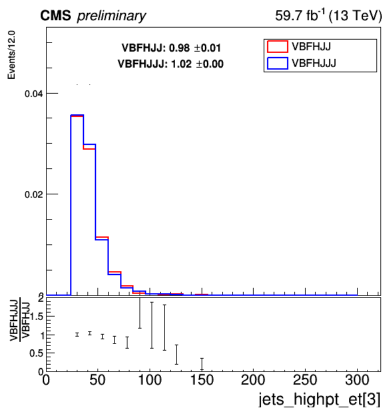
(a)



(b)

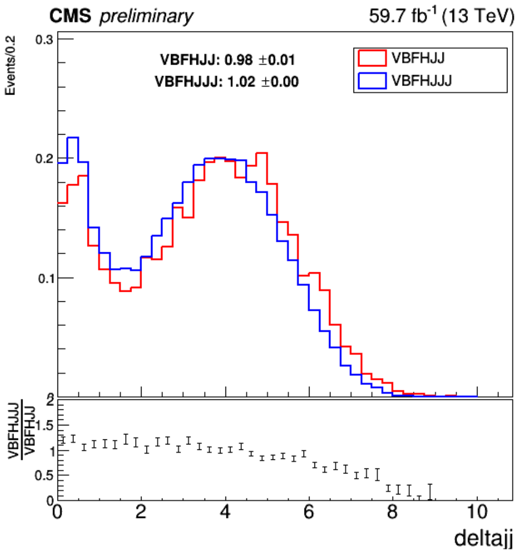


(c)

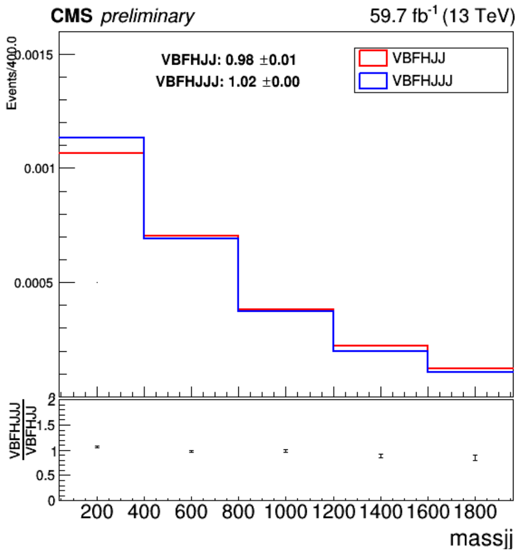


(d)

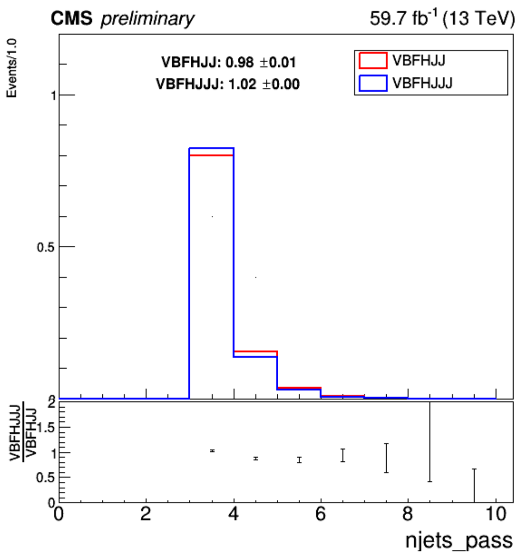
Figure A.1: Comparison plots for the two signal samples VBF-H2J and VBF-H3J for the 3^{rd} jet variables: the transverse momentum of the jet (a), the pseudorapidity of the jet (b), the azimuthal angle of the jet (c), and the transverse energy of the jet (d).



(a)



(b)



(c)

Figure A.2: Comparison plots for the two signal samples VBF-H2J and VBF-H3J for the delta R between the two jets with the highest pT (a), invariant di-jet mass (b), and the number of jets passing the jets selection criteria (c).

Appendix B

Modelling the 3^{rd} jet in the qqZZ sample

B.1 DJR Plots for qqZZ + jets sample

In preparation for LHC Run 3 analysis and to have the correct modelling for the 3^{rd} jet in the backgrounds that affect the VBF Higgs production analysis, the MadGraph event generator has been used to generate the qqZZ process with additional jets up to 3 jets. When one generates the process with additional jets, there is an issue of double-counting of events because of the Parton shower. To avoid this double-counting, we should use the matching scale "qCut", representing the separation between the shower and the matrix element, defined in Pythia cards. However, this qCut is not strictly defined; hence, one needs to scan over it and look for the variable independent of qCut. For instance, to look at the variable “Differential Jet Rate (DJR)”. This observable allows us to measure how smooth the transition is when going from N to $N + 1$ jet configuration. Thus it is quite sensitive to the matching procedure. We choose the value of qCut for which the DJR plot looks smooth.

The plots for the DJR scan to check the MLM merging quality are shown in Figs. B.1 - B.4, where you can see different qCuts have been scanned, 20,27,30 and 35. According to the scan done, the qCut has been selected to be set at 20.

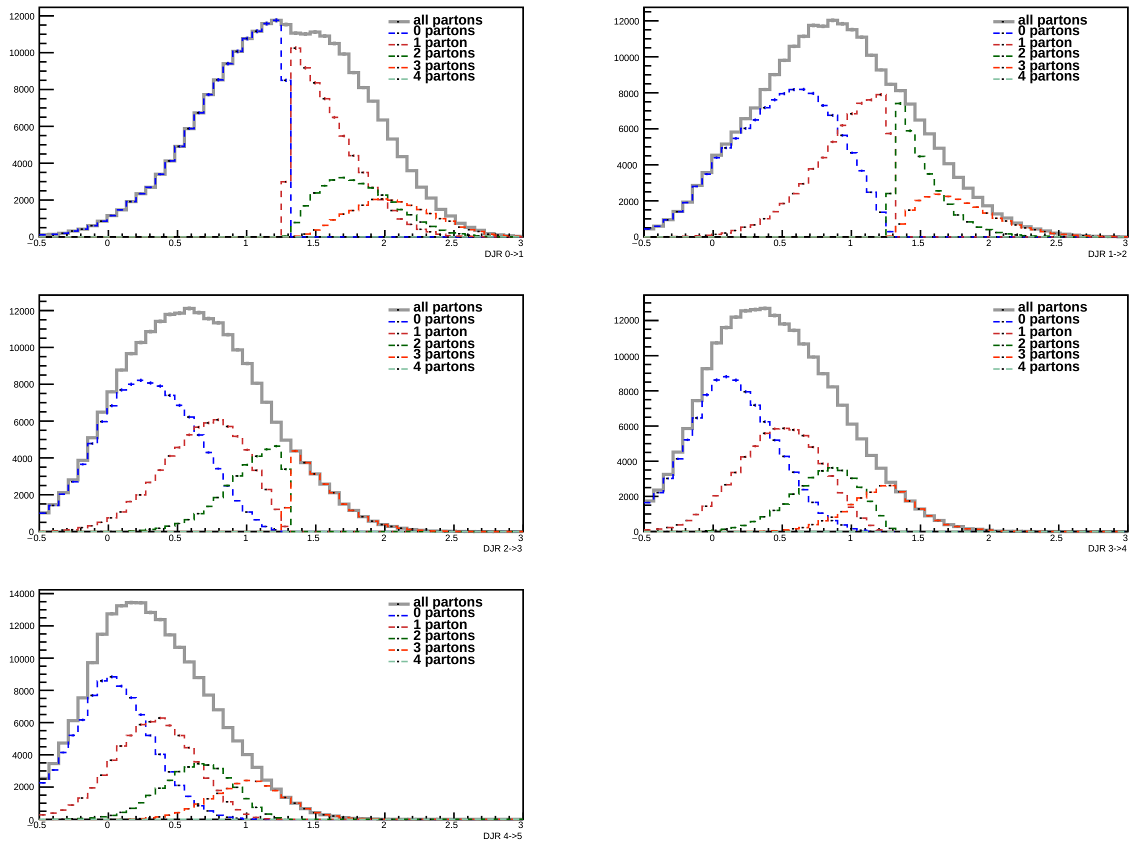


Figure B.1: DJR plots with qCut 20 for the Zll jet binned samples with additional 0,1,2,3 jets using MLM merging.

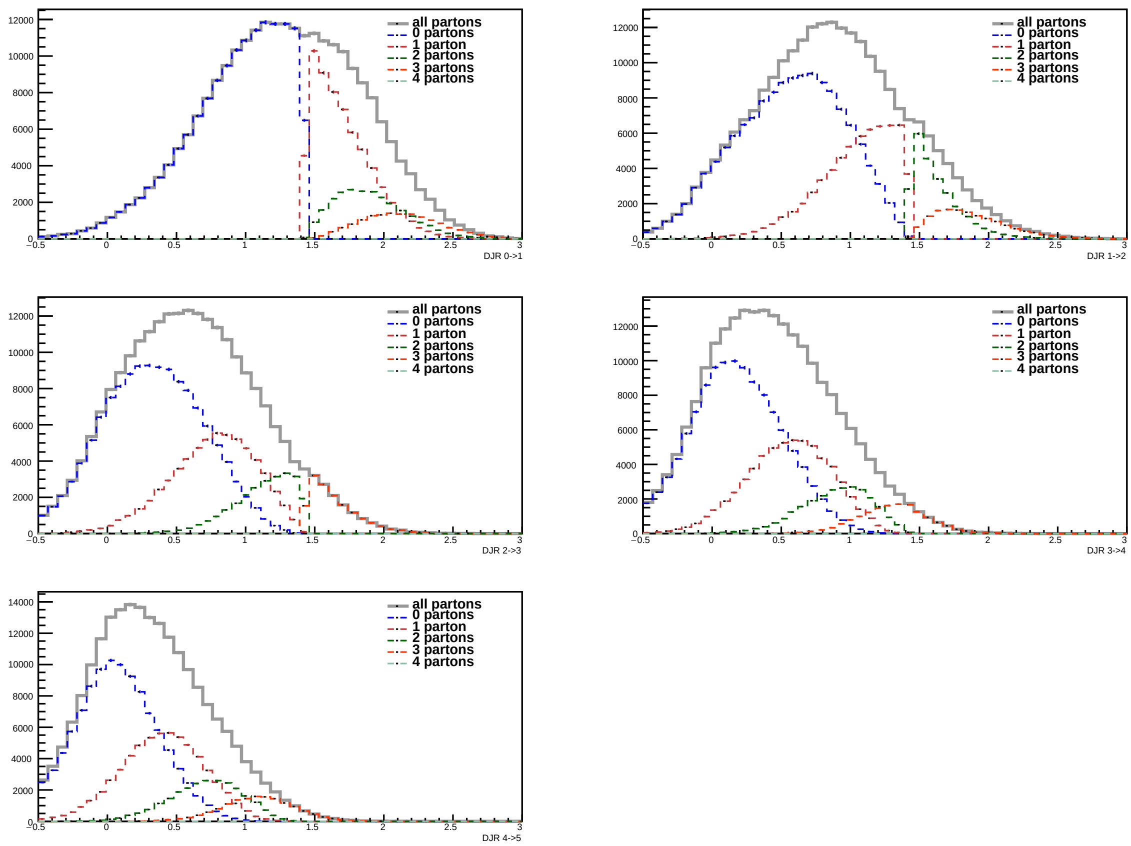


Figure B.2: DJR plots with qCut 27 for the Zll jet binned samples with additional 0,1,2,3 jets using MLM merging.

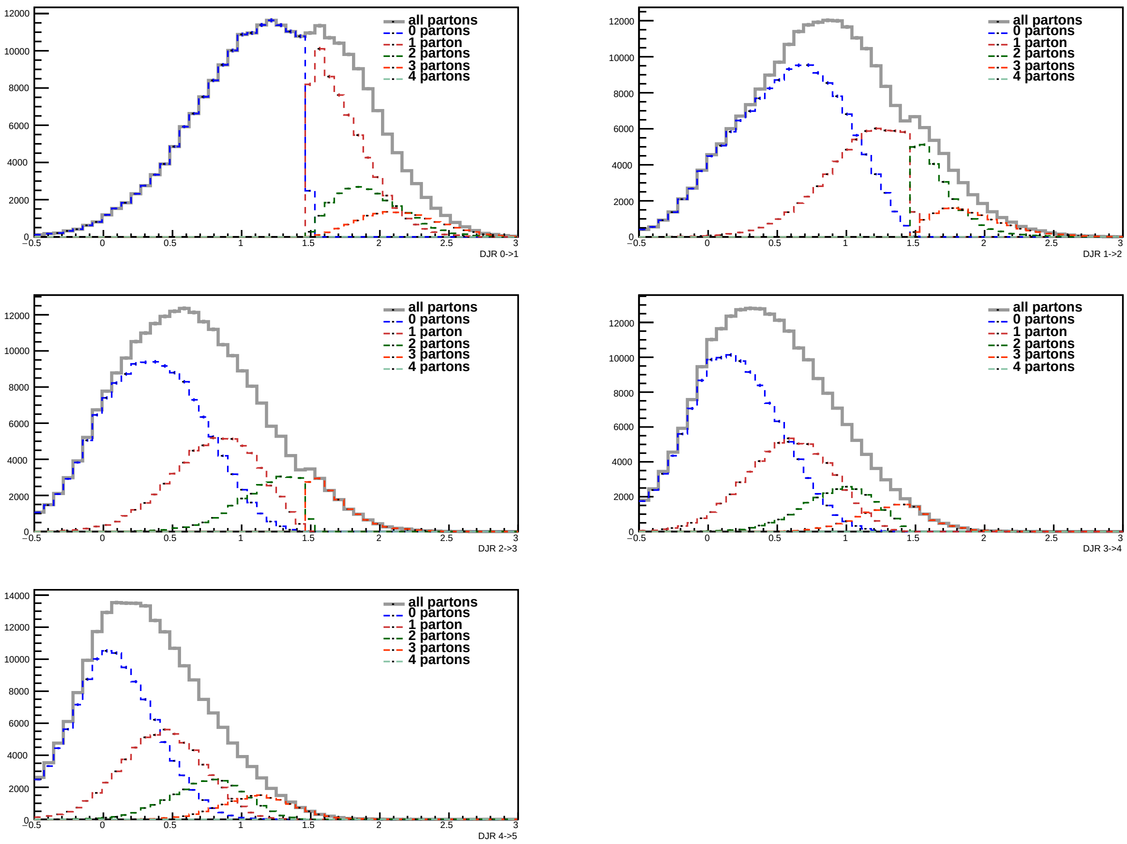


Figure B.3: DJR plots with qCut 30 for the Zll jet binned samples with additional 0,1,2,3 jets using MLM merging.

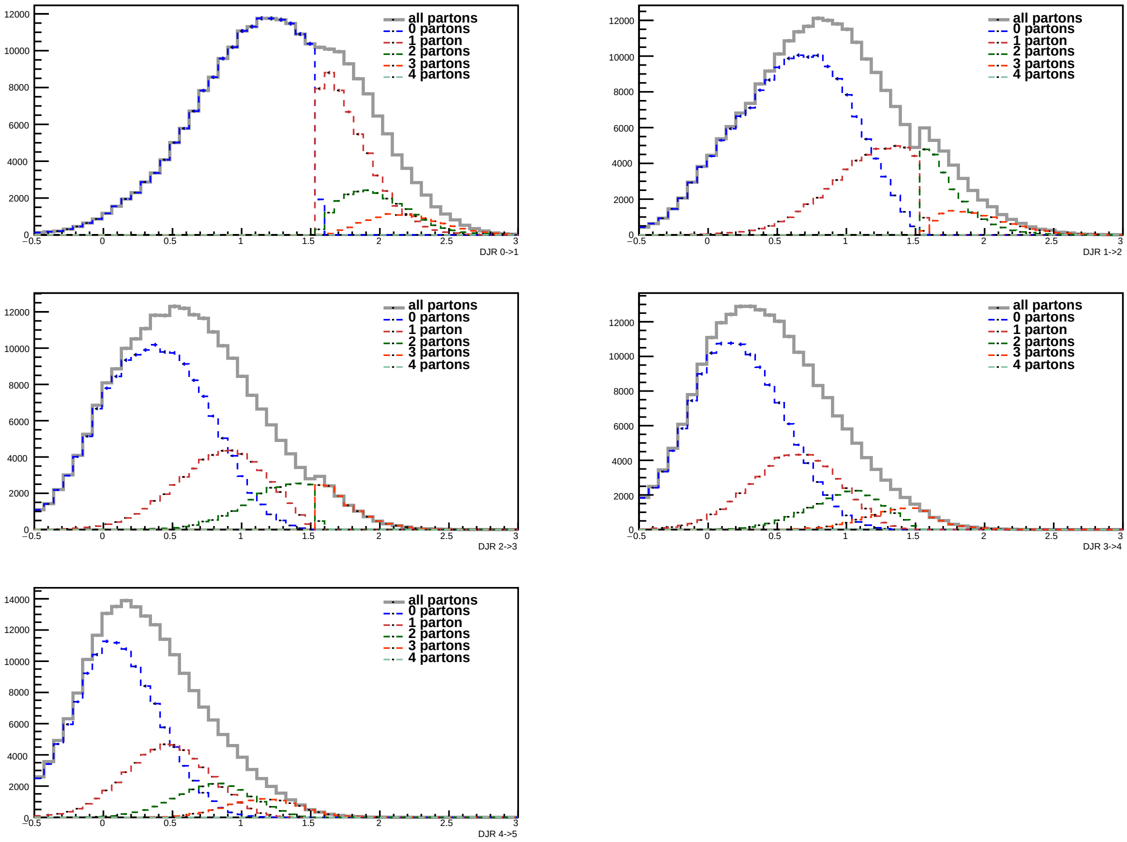


Figure B.4: DJR plots with qCut 35 for the Zll jet binned samples with additional 0,1,2,3 jets using MLM merging.

Acronyms

ALICE Alarge Ion Collider Experiment

ANN Artificial Neural Network

APV Analogue Pipeline Voltage mode

ATLAS A Toroidal LHC ApparatuS

BDT Boosted Decision Trees

BEH Brout-Englert-Higgs

BX Bunch Crossings

CERN Center for European Nuclear Research

CKF Combinatorial Kalman Filter

CMS Compact Muon Solenoid

CP Charge-Parity

CR Control Regions

CSC Cathode Strip Chamber

CTF Combinatorial Track Finder

DIS Deep-Inelastic Scattering

DNN Deep Neural Network

DT Drift Tube

DY Drell-Yan

EB ECAL Barrel

ECAL Electromagnetic Calorimeter

EE ECAL Endap

EW Electro-Weak

FASER ForwArd Search ExpeRiment

FE Front-End

FR Fake Rate

FSR Final State Radiation

GEM Gas Electron Multiplier

GSF Gaussian Sum Filter

HB HCAL Barrel detector

HCAL Hadronic Calorimeter

HE HCAL Endcap detector

HF HCAL Forward detector

HL-LHC High Luminosity Large Hadron Collider

HLT High Level Trigger

HO HCAL Outer calorimeter

HPL High-Pressure Laminate

ID Identification

IP Impact Parameter

ISR Initial State Radiation

JEC Jet Energy Correction

JER Jet Energy Resolution

JES Jet Energy Scale

KF Kalman Filter

L1 Level-1

LHC Large Hadron Collider

LHCb Large Hadron Collider beauty

LHCf LHC-forward

LO Leading Order

LS Long Shutdown

MC Monte Carlo

ME Matrix Element

MoEDAL Monopole and Exotics Detector At the LHC

MVA Multivariate Approach

NDOF Number of Degrees of Freedom

NLO Next-to-Leading Order

NN Neural Network

NNLO Next-to-Next-to-Leading Order

PDF Parton Distribution Functions

PF Particle Flow

PLR Profile Likelihood Ratio

PMNS Pontecorvo-Maki-Nakagawa-Sakata

POI Parameters of Interest

PS Parton Shower

PU Pile-Up

PV Primary Vertex

Q Charge

QCD Quantum Chromodynamics

ReLU Rectified Linear Unit

ROC Receiver Operating Characteristic curve

RPC Resistive Plate Chamber

S Spin

SIP Significance of the Impact Parameter

SM Standard Model

SR Signal Region

SSB Spontaneous Symmetry Breaking

TEC Tracker Endcaps

TIB Tracker Inner Barrel

TID Tracker Inner Disks

TOB Tracker Outer Barrel

TOTEM TOTal Elastic and diffractive cross section Measurement

VBF Vector Boson Fusion

XGBoost eXtreme Gradient Boosting

XS Cross-Section

Bibliography

[1] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". *Phys. Rev. Lett.*, 13:321–323, Aug 1964. URL https://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[2] Peter W. Higgs. "Broken symmetries, massless particles and gauge fields". *Phys. Lett.*, 12: 132–133, 1964. doi: 10.1016/0031-9163(64)91136-9.

[3] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". *Phys. Rev. Lett.*, 13: 508–509, 1964. doi: 10.1103/PhysRevLett.13.508.

[4] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Physics Letters B*, 716(1):1–29, 2012. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2012.08.020. URL https://www.sciencedirect.com/science/article/pii/S037026931200857X.

[5] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". *Physics Letters B*, 716(1):30–61, 2012. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2012.08.021. URL https://www.sciencedirect.com/science/article/pii/S0370269312008581.

[6] Martin Flechl. "Higgs boson discovery and recent results". In *22nd International Workshop on High Energy Physics and Quantum Field Theory*, 10 2015. URL https://arxiv.org/abs/1510.01924.

[7] D. Zeppenfeld, R. Kinnunen, A. Nikitenko, and E. Richter-Was. "Measuring Higgs boson couplings at the LHC". *Phys. Rev. D*, 62:013009/1–10. 25 p, Feb 2000. doi: 10.1103/PhysRevD.62.013009. URL https://cds.cern.ch/record/425540. 25 pages, Revtex, 1 figure Report-no: MADPH-00-1157.

[8] A Purcell. "Standard Model infographic", 2012. URL http://bit.ly/2UKxo1R. accessed: 14.07.2021.

[9] Takaaki Kajita. "Nobel Lecture: Discovery of atmospheric neutrino oscillations". *Rev. Mod. Phys.*, 88:030501, Jul 2016. doi: 10.1103/RevModPhys.88.030501. URL https://link.aps.org/doi/10.1103/RevModPhys.88.030501.

[10] M. Tanabashi and *et al.* "Review of Particle Physics". *Phys. Rev. D*, 98:030001, Aug 2018. doi: 10.1103/PhysRevD.98.030001. URL https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[11] Z Maki, M Nakagawa, and S Sakata. "Remarks on the Unified Model of Elementary Particles". *Progress of Theoretical Physics*, 28(5):870–880, 11 1962. ISSN 0033-068X. doi: 10.1143/PTP.28.870. URL https://doi.org/10.1143/PTP.28.870.

[12] David J. Gross and Frank Wilczek. "Ultraviolet Behavior of Non-Abelian Gauge Theories". *Phys. Rev. Lett.*, 30:1343–1346, Jun 1973. doi: 10.1103/PhysRevLett.30.1343. URL https://link.aps.org/doi/10.1103/PhysRevLett.30.1343.

[13] H. Georgi. *"LIE ALGEBRAS IN PARTICLE PHYSICS. FROM ISOSPIN TO UNIFIED THEORIES"*, volume 54. 1982.

[14] M. D. Schwartz. *"Quantum Field Theory and the Standard Model"*. Cambridge University Press, 2013. doi: {10.1017/9781139540940}.

[15] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. "Global Conservation Laws and Massless Particles". *Phys. Rev. Lett.*, 13:585–587, 1964. URL https://link.aps.org/doi/10.1103/PhysRevLett.13.585.

[16] G.'t Hooft. "Renormalizable Lagrangians for massive Yang-Mills fields". *Nuclear Physics B*, 35(1):167–188, 1971. ISSN 0550-3213. doi: https://doi.org/10.1016/0550-3213(71)90139-8. URL https://www.sciencedirect.com/science/article/pii/0550321371901398.

[17] J Ellis. "Higgs Physics", 2013. URL https://cds.cern.ch/record/1638469.

[18] Y. Nambu and G. Jona-Lasinio. "Dynamical Model of Elementary Particles Based on an Analogy with Superconductivity. 1". *Phys. Rev.*, 122:345–358, Apr 1961. doi: 10.1103/PhysRev.122.345. URL https://link.aps.org/doi/10.1103/PhysRev.122.345.

[19] M. Gell-Mann and M. Lévy. "The axial vector current in beta decay". *Il Nuovo Cimento*, 16:705, 1960. URL https://doi.org/10.1007/BF02859738.

[20] C. Patrignani and *et al.* "Particle Data Group". *Chin. Phys. C*, 40:100001, 2016. URL https://pdg.lbl.gov/2016/reviews/rpp2016-rev-higgs-boson.pdf.

[21] "LHC Higgs Cross Section Working Group". URL https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG.

[22] A.M. Sirunyan and *et al.* "Combined measurements of Higgs boson couplings in proton–proton collisions at $\sqrt{s} = 13\,\mathrm{TeV}$". *The European Physical Journal C*, 79:421, 2019. URL https://doi.org/10.1140/epjc/s10052-019-6909-y.

[23] T. Plehn, D. Rainwater, and D. Zeppenfeld. "Determining the Structure of Higgs Couplings at the CERN Large Hadron Collider". *Phys. Rev. Lett.*, 88:051801, Jan 2002. URL https://link.aps.org/doi/10.1103/PhysRevLett.88.051801.

[24] M Spira. "QCD Effects in Higgs Physics". *Fortschritte der Physik*, 46(3):203–284, 1998. URL https://doi.org/10.1002/(SICI)1521-3978(199804)46:3<203::AID-PROP203>3.0.CO;2-4.

[25] T. Han, G. Valencia, and S. Willenbrock. "Structure-function approach to vector-boson scattering in pp collisions". *Phys. Rev. Lett.*, 69:3274–3277, Dec 1992. URL https://link.aps.org/doi/10.1103/PhysRevLett.69.3274.

[26] R. V. Harlander, J. Vollinga, and M. M. Weber. "Gluon-induced weak boson fusion". *Phys. Rev. D*, 77:053010, Mar 2008. URL https://link.aps.org/doi/10.1103/PhysRevD.77.053010.

[27] P. Bolzoni, F. Maltoni, S. Moch, and M. Zaro. "Higgs Boson Production via Vector-Boson Fusion at Next-to-Next-to-Leading Order in QCD". *Phys. Rev. Lett.*, 105:011801, Jul 2010. URL https://link.aps.org/doi/10.1103/PhysRevLett.105.011801.

[28] The ATLAS Collaboration. "Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector". *Phys. Rev. D*, 90:112015, Dec 2014. URL https://link.aps.org/doi/10.1103/PhysRevD.90.112015.

[29] The CMS collaboration. "Observation of the diphoton decay of the Higgs boson and measurement of its properties". *Eur. Phys. J. C*, 74, 2014. URL https://doi.org/10.1140/epjc/s10052-014-3076-z.

[30] The ATLAS Collaboration. "Measurements of Higgs boson production and couplings in the four-lepton channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector". *Phys. Rev. D*, 91:012006, Jan 2015. URL https://link.aps.org/doi/10.1103/PhysRevD.91.012006.

[31] The CMS Collaboration. "Measurement of the properties of a Higgs boson in the four-lepton final state". *Phys. Rev. D*, 89:092007, May 2014. URL https://link.aps.org/doi/10.1103/PhysRevD.89.092007.

[32] The ATLAS Collaboration. "Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector". *Phys. Rev. D*, 92:012006, Jul 2015. URL https://link.aps.org/doi/10.1103/PhysRevD.92.012006.

[33] The CMS collaboration. "Measurement of Higgs boson production and properties in the WW decay channel with leptonic final states". *J. High Energ. Phys.*, 96, 2014. URL https://doi.org/10.1007/JHEP01(2014)096.

[34] The CMS Collaboration. "Study of the Mass and Spin-Parity of the Higgs Boson Candidate via Its Decays to Z Boson Pairs". *Phys. Rev. Lett.*, 110:081803, Feb 2013. URL https://link.aps.org/doi/10.1103/PhysRevLett.110.081803.

[35] The ATLAS Collaboration. "Evidence for the spin-0 nature of the Higgs boson using ATLAS data". *Physics Letters B*, 726(1):120–144, 2013. ISSN 0370-2693. URL https://www.sciencedirect.com/science/article/pii/S0370269313006527.

[36] G. Aad, B. Abbott, and *et al.* "Higgs boson production cross-section measurements and their EFT interpretation in the 4ℓ decay channel at $\sqrt{s}$ =13 TeV with the ATLAS detector".

European Physical Journal. C, Particles and Fields, 80(10), 10 2020. URL http://dx.doi.org/10.1140/epjc/s10052-020-8227-9.

[37] A. M. Sirunyan, A. Tumasyan, and *et al.* "Measurements of production cross sections of the Higgs boson in the four-lepton final state in proton–proton collisions at $\sqrt{s} = 13\,\text{TeV}$". 81, 2021. URL https://doi.org/10.1140/epjc/s10052-021-09200-x.

[38] A.M. Sirunyan, A. Tumasyan, and *et al.* "A measurement of the Higgs boson mass in the diphoton decay channel". *Physics Letters B*, 805:135425, 2020. ISSN 0370-2693. doi: https://doi.org/10.1016/j.physletb.2020.135425. URL https://www.sciencedirect.com/science/article/pii/S037026932030229X.

[39] Jehad Mousa, A Tumasyan, and *et al.* "Measurements of $t\bar{t}$H Production and the CP Structure of the Yukawa Interaction between the Higgs Boson and Top Quark in the Diphoton Decay Channel". *Physical Review Letters*, 125, 08 2020. URL https://link.aps.org/doi/10.1103/PhysRevLett.125.061801.

[40] G Aad, B Abbott, and *et al.* "Measurements of the Higgs boson inclusive and differential fiducial cross sections in the 4ℓ decay channel at $\sqrt{s}$ =13 TeV". *The European Physical Journal C*, 80, 10 2020. URL https://doi.org/10.1140/epjc/s10052-020-8223-0.

[41] The ATLAS Collaboration. "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV". *J. High Energ. Phys.*, 2016(8), Aug 2016. ISSN 1029-8479. URL http://dx.doi.org/10.1007/JHEP08(2016)045.

[42] The ATLAS Collaboration. "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments". *Phys. Rev. Lett.*, 114:191803, May 2015. URL https://link.aps.org/doi/10.1103/PhysRevLett.114.191803.

[43] The CMS Collaboration. "Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector". *Physics Letters B*, 779:283–316, 2018. ISSN 0370-2693. URL https://doi.org/10.1016/j.physletb.2018.02.004.

[44] The ATLAS Collaboration. "Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". *Phys. Rev. D*, 99:072001, Apr 2019. URL https://link.aps.org/doi/10.1103/PhysRevD.99.072001.

[45] The CMS Collaboration. "Observation of $t\bar{t}H$ Production". *Phys. Rev. Lett.*, 120:231801, Jun 2018. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.231801.

[46] The ATLAS Collaboration. "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector". *Physics Letters B*, 784:173–191, 2018. ISSN 0370-2693. URL https://www.sciencedirect.com/science/article/pii/S0370269318305732.

[47] The CMS Collaboration. "Evidence for the Higgs boson decay to a bottom quark–antiquark pair". *Physics Letters B*, 780:501–532, 2018. ISSN 0370-2693. URL https://www.sciencedirect.com/science/article/pii/S0370269318301618.

[48] The ATLAS Collaboration. "Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector". *Physics Letters B*, 786:59–86, 2018. ISSN 0370-2693. URL https://www.sciencedirect.com/science/article/pii/S0370269318307056.

[49] The ATLAS Collaboration. "A search for the dimuon decay of the Standard Model Higgs boson with the ATLAS detector". *Physics Letters B*, 812:135980, 2021. ISSN 0370-2693. URL https://www.sciencedirect.com/science/article/pii/S0370269320307838.

[50] The CMS Collaboration. "Evidence for Higgs boson decay to a pair of muons". *J. High Energ. Phys.*, 148, 2021. URL https://doi.org/10.1007/JHEP01(2021)148.

[51] The ATLAS Collaboration. "Measurement of the Higgs boson mass in the $H \to ZZ^* \to 4\ell$ decay channel with $\sqrt{s}$ = 13 TeV pp collisions using the ATLAS detector at the LHC". Technical report, CERN, Geneva, Apr 2020. URL https://cds.cern.ch/record/2714883.

[52] The CMS Collaboration. "Measurements of the Higgs boson width and anomalous HVV couplings from on-shell and off-shell production in the four-lepton final state". *Phys. Rev. D*, 99:112003, Jun 2019. doi: 10.1103/PhysRevD.99.112003. URL https://link.aps.org/doi/10.1103/PhysRevD.99.112003.

[53] Lyndon Evans and Philip Bryant. "LHC Machine". *JINST*, 3(08):S08001, aug 2008. URL https://doi.org/10.1088/1748-0221/3/08/s08001.

[54] HiLumi Collaboration. "The HL-LHC project". URL https://hilumilhc.web.cern.ch/content/hl-lhc-project.

[55] E. Mobs. "The CERN accelerator complex - August 2018". Aug 2018. URL https://cds.cern.ch/record/2636343. General Photo.

[56] The ATLAS Collaboration and *et al.* "The ATLAS Experiment at the CERN Large Hadron Collider". *JINST*, 3(08):S08003, aug 2008. URL https://doi.org/10.1088/1748-0221/3/08/s08003.

[57] The CMS Collaboration. "The CMS experiment at the CERN LHC". *JINST*, 3:S08004, aug 2008. URL https://doi.org/10.1088/1748-0221/3/08/s08004.

[58] The LHCb Collaboration and *et al.* "The LHCb Detector at the LHC". *JINST*, 3(08):S08005, aug 2008. URL https://doi.org/10.1088/1748-0221/3/08/s08005.

[59] The ALICE Collaboration and *et al.* "The ALICE experiment at the CERN LHC". *JINST*, 3(08):S08002, aug 2008. URL https://doi.org/10.1088/1748-0221/3/08/s08002.

[60] CMS Luminosity public plots. URL https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[61] I. Neutelings. "CMS Wiki Pages, How to draw diagrams in LaTeX with TikZ". URL https://wiki.physik.uzh.ch/cms/latex:tikz.

[62] The CMS Collaboration. "CMS, tracker technical design report". 1998.

[63] The CMS Collaboration. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". *JINST*, 9(10):P10009, oct 2014. URL https://doi.org/10.1088/1748-0221/9/10/p10009.

[64] Hannsjörg Weber. "The phase-1 upgrade of the CMS pixel detector". In *2016 IEEE Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD)*, pages 1–4, 2016. URL https://doi.org/10.1109/NSSMIC.2016.8069719.

[65] K. Dilsiz, S. DURGUT, K. Yİ, and L. SPIEGEL. "Quality control of silicon pixel wafers for the CMS Phase-1 pixel upgrade". *TURKISH JOURNAL OF PHYSICS*, 43:541–550, 12 2019. URL https://doi.org/10.3906/fiz-1907-2.

[66] The CMS Collaboration. *"The CMS electromagnetic calorimeter project: Technical Design Report"*. Technical design report. CMS. CERN, Geneva, 1997. URL https://cds.cern.ch/record/349375.

[67] M. Cipriani. "Performance of the CMS electromagnetic calorimeter during the LHC Run II and its role in precision physics measurements". *PoS*, LHCP2018:084, 2018. URL https://doi.org/10.22323/1.321.0084.

[68] The CMS Collaboration. *"The CMS hadron calorimeter project: Technical Design Report"*. Technical design report. CMS. CERN, Geneva, 1997. URL http://cds.cern.ch/record/357153.

[69] A.M. Sirunyan and *et al.* "Calibration of the CMS hadron calorimeters using proton-proton collision data at $\sqrt{s}$ =13 TeV". *JINST*, 15(05):P05002, may 2020. URL https://doi.org/10.1088/1748-0221/15/05/p05002.

[70] S. Abdullin et al. "The CMS barrel calorimeter response to particle beams from 2 GeV/c to 350 GeV/c". *Eur. Phys. J. C*, 60:359–373, 2009. URL https://doi.org/10.1140/epjc/s10052-009-0959-5.

[71] J. Man and et al. "CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter". 9 2012. URL https://www.osti.gov/biblio/1151651.

[72] J. G. Layter. *"The CMS muon project: Technical Design Report"*. Technical design report. CMS. CERN, Geneva, 1997. URL https://cds.cern.ch/record/343814.

[73] A.M. Sirunyan and *et al.* "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s}$ =13 TeV". *JINST*, 13(06):P06015, jun 2018. URL https://doi.org/10.1088/1748-0221/13/06/p06015.

[74] D. Abbaneo, M. Abbrescia, P. Aspell, S. Bianco, K. Hoepfner, M. Hohlmann, M. Maggi, G. De Lentdecker, A. Safonov, A. Sharma, and M. Tytgat. “A GEM Detector System for an Upgrade of the High-eta Muon Endcap Stations GE1/1 + ME1/1 in CMS”, 2012. URL https://arxiv.org/abs/1211.1494.

[75] W. Ahmed and *et al.* “The Triple GEM Detector Control System for CMS forward muon spectrometer upgrade”. *JINST*, 12(02):P02003–P02003, feb 2017. URL https://doi.org/10.1088/1748-0221/12/02/p02003.

[76] G L Bayatyan and *et al. “CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems”*. Technical design report. CMS. URL https://cds.cern.ch/record/706847.

[77] The CMS Collaboration. “The CMS high level trigger”. *Eur. Phys. J. C*, 46:605–667, 2006. URL https://doi.org/10.1140/epjc/s2006-02495-8.

[78] A.M. Sirunyan and *et al.* “Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s}$ =13 TeV”. *JINST*, 15(10):P10017, oct 2020. URL https://doi.org/10.1088/1748-0221/15/10/p10017.

[79] A Tapper and Darin Acosta. “CMS Technical Design Report for the Level-1 Trigger Upgrade”. Technical report, Jun 2013. URL https://cds.cern.ch/record/1556311.

[80] A. Zabi, , and *et al.* “The CMS Level-1 Calorimeter Trigger for the LHC Run II”. *Journal of Instrumentation*, 12(01):C01065–C01065, jan 2017. URL https://doi.org/10.1088/1748-0221/12/01/c01065.

[81] CMS Collaboration. “Particle-flow reconstruction and global event description with the CMS detector”. *JINST*, 12:P10003. 82 p, Jun 2017. URL https://cds.cern.ch/record/2270046.

[82] R. Frühwirth. “Application of Kalman filtering to track and vertex fitting”. 262(2):444–450, 1987. ISSN 0168-9002. URL https://www.sciencedirect.com/science/article/pii/0168900287908874.

[83] CMS Collaboration. “Muon reconstruction performance during Run II”. Jul 2019. URL https://cds.cern.ch/record/2682902.

[84] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov. “Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC”. *Journal of Physics G: Nuclear and Particle Physics*, 31(9):N9–N20, jul 2005. URL https://doi.org/10.1088/0954-3899/31/9/n01.

[85] CMS Collaboration. “Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s}$ = 8 TeV. Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at sqrt(s)=8 TeV”. *JINST*, 10:P06005. 63 p, Feb 2015. URL https://cds.cern.ch/record/1988091.

[86] “Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s}$ = 8 TeV”. *JINST*, 10(08):P08010–P08010, aug 2015. URL https://doi.org/10.1088/1748-0221/10/08/p08010.

[87] M. Cacciari, G. P. Salam, and G. Soyez. “The anti-k$_t$ jet clustering algorithm”. *J. High Energ. Phys.*, 2008(04):063–063, apr 2008. URL https://doi.org/10.1088/1126-6708/2008/04/063.

[88] CMS Collaboration. “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”. *JINST*, 13(05):P05011–P05011, may 2018. URL https://doi.org/10.1088/1748-0221/13/05/p05011.

[89] E.D. Malaza. “Multiplicity distributions in quark and gluon jets to O(α_s)”. *Z. Phys. C - Particles and Fields*, 31:143–150, 1986. URL https://doi.org/10.1007/BF01559605.

[90] “Jet algorithms performance in 13 TeV data”. Technical report, CERN, Geneva, 2017. URL https://cds.cern.ch/record/2256875.

[91] URL https://theory.slac.stanford.edu/our-research/simulations.

[92] R. Keith Ellis, W. James Stirling, and B. R. Webber. *“QCD and collider physics”*, volume 8. Cambridge University Press, 2 2011. ISBN 978-0-511-82328-2, 978-0-521-54589-1. URL https://inspirehep.net/literature/328604.

[93] A. Buckley and *et al.* “MCnet-11-01 General-purpose event generators for LHC physics”. *Physics Reports*, 504, 01 2011. URL https://doi.org/10.1016/j.physrep.2011.03.005.

[94] Bo Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand. “Parton Fragmentation and String Dynamics”. *Phys. Rept.*, 97:31–145, 1983. URL https://doi.org/10.1016/0370-1573(83)90080-7.

[95] Nick Van Remortel. “Minimum bias and underlying event studies at CMS”. Berlin, Germany, 2009. Science Wise Publ. URL http://cds.cern.ch/record/1199942.

[96] Torbjörn Sjöstrand and *et al.* “An introduction to PYTHIA 8.2”. *Computer Physics Communications*, 191:159–177, 2015. ISSN 0010-4655. URL https://www.sciencedirect.com/science/article/pii/S0010465515000442.

[97] J. Alwall, R. Frederix, S. Frixione, and *et al.* “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations.”. *J. High Energ. Phys.*, 79, 2014. URL https://www.sciencedirect.com/science/article/pii/S0010465508000441.

[98] C. Oleari. “The POWHEG BOX”. *Nuclear Physics B - Proceedings Supplements*, 205-206: 36–41, 2010. ISSN 0920-5632. URL https://www.sciencedirect.com/science/article/pii/S0920563210001994.

[99] S. Agostinelli and *et al.* “Geant4—a simulation toolkit”. *In: Nucl. Instrum. Meth. A*, 506(3): 250–303, 2003. ISSN 0168-9002. URL https://www.sciencedirect.com/science/article/pii/S0168900203013688.

[100] Walaa Elmetenawee. “CMS track reconstruction performance during Run 2 and developments for Run 3”. *PoS*, ICHEP2020:733, 2021. URL https://doi.org/10.22323/1.390.0733.

[101] Mia Tosi. "Tracking at High Level Trigger in CMS". *Nucl. Part. Phys. Proc.*, 273-275:2494–2496, 2016. doi: https://doi.org/10.1016/j.nuclphysbps.2015.09.436.

[102] The CMS Collaboration. "2017 tracking performance plots", Apr 2017. URL https://cds.cern.ch/record/2290524.

[103] M.J. French and *et al.* "Design and results from the APV25, a deep sub-micron CMOS front-end chip for the CMS tracker". *Nucl. Instrum. Meth.*, 466(2):359–365, 2001. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(01)00589-7. URL https://www.sciencedirect.com/science/article/pii/S0168900201005897.

[104] N. Adam, J. Berryhill, V. Halyo, A. Hunt, and K. Mishra. "Generic Tag and Probe Tool for Measuring Efficiency at CMS with Early Data". URL https://cms.cern.ch/iCMS/user/noteinfo?cmsnoteid=CMS%20AN-2009/111. CMS Note 2009/111, 2009.

[105] G. Petrucciani. "Measurement of the tracking efficiency for muons using the tag and probe method". URL https://cms.cern.ch/iCMS/user/noteinfo?cmsnoteid=CMS%20AN-2010/149. CMS Note 2010/149, 2010.

[106] Walaa Elmetenawee for the CMS Collaboration. "Muon tracking performance in the CMS Run-2 Legacy data using the tag-and-probe technique". Jul 2020. URL https://cds.cern.ch/record/2724492.

[107] W. Elmetenawee and *et al.* "Effects of the electronic threshold on the performance of the RPC system of the CMS experiment". *JINST*, 15:C09025. 7 p, May 2020. URL https://doi.org/10.1088/1748-0221/15/09/C09025.

[108] M. Abbrescia, P. Fonte, and V. Peskov. *"Resistive Gaseous Detectors: designs, performance, and perspective"*. Wiley-VCH, April 2018. ISBN 978-3-527-34076-7, 978-0-521-54589-1. URL https://doi.org/10.1002/9783527698691.

[109] K. Bunkowski and J. Krolikowski. "Optimization, Synchronization, Calibration and Diagnostic of the RPC PAC Muon Trigger System for the CMS detector". 01 2009.

[110] The CMS Collaboration. "RPC Detector Performance Results for 2016 and 2017", Feb 2018. URL https://cds.cern.ch/record/2305542.

[111] M. Abbrescia and *et al.* "Properties of C2H2F4-based gas mixture for avalanche mode operation of resistive plate chambers". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 398(2):173–179, 1997. ISSN 0168-9002. URL https://www.sciencedirect.com/science/article/pii/S016890029700689X.

[112] P. Camarri and *et al.* "Streamer suppression with SF6 in RPCs operated in avalanche mode". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 414(2):317–324, 1998. ISSN 0168-9002. URL https://www.sciencedirect.com/science/article/pii/S0168900298005762.

[113] E. C. Zeballos and *et al.* "Effect of adding SF6 to the gas mixture in a multigap resistive plate chamber". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 419(2):475–478, 1998. ISSN 0168-9002. URL https://www.sciencedirect.com/science/article/pii/S0168900298008237.

[114] "RPC Performance Results for 2016", 2017. URL https://twiki.cern.ch/twiki/bin/view/CMSPublic/RPCDPGResultsforVenice.

[115] S. Park and *et al.* "Production of gas gaps for the Forward RPCs of the CMS experiment". *Nucl. Instrum. Meth. A*, 550:551–558, 2005. URL https://doi.org/10.1016/j.nima.2005.05.052.

[116] M. Abbrescia and *et al.* "New developments on front end electronics for the CMS Resistive Plate Chambers". *Nucl. Instrum. Meth. A*, 456:143–149, 2000. URL https://doi.org/10.1016/S0168-9002(00)00980-3.

[117] The CMS collaboration. "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$TeV". *JINST*, 7(10):P10002–P10002, oct 2012. URL https://doi.org/10.1088/1748-0221/7/10/p10002.

[118] Jan Eysermans and María Isabel Pedraza Morales and. "Operation and performance of the CMS Resistive Plate Chambers during LHC run II". *Journal of Physics: Conference Series*, 912:012015, oct 2017. URL https://doi.org/10.1088/1742-6596/912/1/012015.

[119] The CMS Collaboration. "CMS Luminosity Measurements for the 2016 Data Taking Period". Technical report, CERN, Geneva, 2017. URL https://cds.cern.ch/record/2257069.

[120] The CMS Collaboration. "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV". Technical report, CERN, Geneva, 2018. URL https://cds.cern.ch/record/2621960.

[121] The CMS Collaboration. "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV". Technical report, CERN, Geneva, 2019. URL https://cds.cern.ch/record/2676164.

[122] M. Ahmad and *et al.* for the CMS Collaboration. "Measurements of properties of the Higgs boson in the four-lepton final state at $\sqrt{s} = 13$ TeV with full Run II data", 2019. CMS AN-2019/139.

[123] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. "NLO vector-boson production matched with shower in POWHEG". *J. High Energ. Phys.*, 2008(07):060–060, jul 2008. URL https://doi.org/10.1088/1126-6708/2008/07/060.

[124] Paolo Nason. "A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms". *J. High Energ. Phys.*, 2004(11):040–040, nov 2004. URL https://doi.org/10.1088/1126-6708/2004/11/040.

[125] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". *J. High Energ. Phys.*, 2007(11):070–070, nov 2007. URL https://doi.org/10.1088/1126-6708/2007/11/070.

[126] Bagnaschi, E. *et al.* "Higgs production via gluon fusion in the POWHEG approach in the SM and in the MSSM". *J. High Energ. Phys.*, 02:088, 2012. URL https://doi.org/10.1007/JHEP02(2012)088.

[127] P. Nason and C. Oleari. "NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG". *J. High Energ. Phys.*, 02:037, 2010. URL https://doi.org/10.1007/JHEP02(2010)037.

[128] Luisoni, G. and *et al.* "$HW^{\pm}/HZ + 0$ and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO". *J. High Energ. Phys.*, 10:083, 2013. URL https://doi.org/10.1007/JHEP10(2013)083.

[129] H. B. Hartanto, B. Jäger, L. Reina, and D. Wackeroth. "Higgs boson production in association with top quarks in the POWHEG BOX". *Phys. Rev. D*, 91:094003, May 2015. URL https://link.aps.org/doi/10.1103/PhysRevD.91.094003.

[130] Yanyan Gao, Andrei V. Gritsan, Zijin Guo, Kirill Melnikov, Markus Schulze, and Nhan V. Tran. "Spin determination of single-produced resonances at hadron colliders". *Phys. Rev. D*, 81:075022, Apr 2010. URL https://link.aps.org/doi/10.1103/PhysRevD.81.075022.

[131] D. de Florian et al. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". 2/2017, 10 2016. URL https://doi.org/10.23731/CYRM-2017-002.

[132] P. Nason and G. Zanderighi. "W^+W^-, WZ, ZZ production in the POWHEG-BOX-V2". *European Physics Journal C*, 74:2702, 2014. URL https://doi.org/10.1140/epjc/s10052-013-2702-5.

[133] John M. Campbell and R.K. Ellis. "MCFM for the Tevatron and the LHC, journal = Nuclear Physics B - Proceedings Supplements". 205-206:10–15, 2010. ISSN 0920-5632. URL https://www.sciencedirect.com/science/article/pii/S0920563210001945. Loops and Legs in Quantum Field Theory.

[134] J. M. Campbell, R. K. Ellis, and C. Williams. "Bounding the Higgs width at the LHC using full analytic results for $gg \to e^-e^+\mu^-\mu^+$". *J. High Energ. Phys.*, 04:060, 2014. URL https://doi.org/10.1007/JHEP04(2014)060.

[135] J. Allison and *et al.* "Geant4 developments and applications". *IEEE Transactions on Nuclear Science*, 53(1):270–278, 2006. URL https://doi.org/10.1109/TNS.2006.869826.

[136] B. Jäger, F. Schissler, D. Zeppenfeld. "Parton-shower effects on Higgs boson production via vector-boson fusion in association with three jets.". *JHEP*, 07:125, 2014. URL https://doi.org/10.1007/JHEP07(2014)125.

[137] T. Figy, V. Hankele, D. Zeppenfeld. "Next-to-leading order QCD corrections to Higgs plus three jet production in vector-boson fusion.". *JHEP*, 02:076, 2008. URL https://doi.org/10.1088/1126-6708/2008/02/076.

[138] K. Arnold *et al.* "VBFNLO: a parton level Monte Carlo for processes with electroweak bosons.". *Comput. Phys. Commun.*, 180:1661–1670, 2009. URL https://doi.org/10.1016/j.cpc.2009.03.006.

[139] J Baglio and *et al.* "VBFNLO: A Parton Level Monte Carlo for Processes with Electroweak Bosons – Manual for Version 2.7.0". July 2011. URL https://arxiv.org/abs/1107.4038.

[140] Jeppe R. Andersen and Jennifer M. Smillie. "QCD and electroweak interference in Higgs production by gauge boson fusion". *Phys. Rev. D*, 75:037301, Feb 2007. URL https://link.aps.org/doi/10.1103/PhysRevD.75.037301.

[141] M. Ciccolini, A. Denner, and S. Dittmaier. "Strong and Electroweak Corrections to the Production of a Higgs Boson$+2$ Jets via Weak Interactions at the Large Hadron Collider". *Phys. Rev. Lett.*, 99:161803, Oct 2007. URL https://link.aps.org/doi/10.1103/PhysRevLett.99.161803.

[142] Terrance Figy, Vera Hankele, and Dieter Zeppenfeld. "Dominant next-to-leading order QCD corrections to Higgs plus three jet production in vector-boson fusion". *J. High Energ. Phys.*, 2008(02):076–076, feb 2008. URL https://doi.org/10.1088/1126-6708/2008/02/076.

[143] The CMS Collaboration. "Performance of Electron and Photon reconstruction in the 2017 Legacy dataset of the CMS Experiment". May 2020. URL https://cds.cern.ch/record/2718815.

[144] T. Chen and C. Guestrin. "XGBoost: A scalable tree boosting system". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016. URL http://dx.doi.org/10.1145/2939672.2939785.

[145] The CMS Collaboration. "Electron and Photon performance in CMS with the full 2017 data sample and additional 2016 highlights for the CALOR 2018 Conference". May 2018. URL https://cds.cern.ch/record/2320638.

[146] M. Ahmad and *et al.* for the CMS Collaboration. "Measurements of properties of the higgs boson in the four-lepton final state at $\sqrt{s} = 13$ TeV", 2015. CMS AN-15/277.

[147] M. Ahmad and *et al.* for the CMS Collaboration. "New results on the study of higgs boson production in the four-lepton final state at $\sqrt{s} = 13$ TeV", 2016. CMS AN-16/217.

[148] Giovanni Petrucciani. "Observation of a new state in the search for the Higgs boson at CMS. The search for the Higgs boson at CMS", Sep 2013. URL https://cds.cern.ch/record/1547904. Presented 22 Jan 2013.

[149] The CMS Collaboration. "Pileup mitigation at CMS in 13 TeV data". *JINST*, 15(09):P09018–P09018, sep 2020. URL https://doi.org/10.1088/1748-0221/15/09/p09018.

[150] The CMS Collaboration. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". *Journal of Instrumentation*, 13(05):P05011–P05011, may 2018. URL https://doi.org/10.1088/1748-0221/13/05/p05011.

[151] M. Bonvini, F. Caola, S. Forte, K. Melnikov, and G. Ridolfi. "Signal-background interference effects for $gg \to H \to W^+W^-$ beyond leading order". *Phys. Rev. D*, 88:034032, Aug 2013. URL https://link.aps.org/doi/10.1103/PhysRevD.88.034032.

[152] C. Li, H. Li, D. Shao, and J. Wang. "Soft gluon resummation in the signal-background interference process of gg(→ h∗) → ZZ". *Journal of High Energy Physics*, 2015, 04 2015. URL https://doi.org/10.1007/JHEP08(2015)065.

[153] K. Gurney. *"An Introduction to Neural Networks"*. UCL Press, 2004. ISBN 1-85728-503-4.

[154] S.S. Haykin. *"Neural Networks and Learning Machines"*. Pearson International Edition. Pearson, 3rd edition, 2009. ISBN 9780131293762. URL https://books.google.it/books?id=KCwWOAAAACAAJ.

[155] URL https://keras.io/. Keras Web Page.

[156] Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. "On the Number of Linear Regions of Deep Neural Networks". In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/109d2dd3608f669ca17920c511c2a41e-Paper.pdf.

[157] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization", 2017.

[158] Matthew D. Zeiler. "ADADELTA: An Adaptive Learning Rate Method", 2012.

[159] N. Srivastava, Geoffrey Hinton, Alex K., Ilya S., and R. Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014. ISSN 1532-4435.

[160] Michiel Botje, Jon Butterworth, Amanda Cooper-Sarkar, Albert de Roeck, Joel Feltesse, Stefano Forte, Alexander Glazov, Joey Huston, Ronan McNulty, Torbjorn Sjostrand, and Robert Thorne. "The PDF4LHC Working Group Interim Recommendations", 2011.

[161] J. C. Collins, D. E. Soper, and G. F. Sterman. "Factorization of Hard Processes in QCD". *Adv. Ser. Direct. High Energy Phys.*, 1989:1–91, 4 2015. URL https://arxiv.org/abs/hep-ph/0409313.

[162] CMS Collaboration. "Search for the standard model Higgs boson in the decay channel $H \to ZZ \to 4l$ in pp collisions". CMS Analysis Note CMS AN-12-141, CERN, 2012.

[163] CMS Collaboration. "Search for associated production of Higgs bosons and top quarks in multilepton final states at $\sqrt{s} = 13$ TeV". CMS Physics Analysis Summary CMS-PAS-HIG-16-022, CERN, 2016.

[164] ATLAS Collaboration. "Evidence for the $H \to b\bar{b}$ decay with the ATLAS detector". Technical Report ATLAS-CONF-2017-041, CERN, 2017.

[165] G. Cowan, K. Cranmer, E. Gross, and *et al.* "Asymptotic formulae for likelihood-based tests of new physics". *Eur. Phys. J. C*, 71:1554, 2011. URL https://doi.org/10.1140/epjc/s10052-011-1554-0.

[166] R. D. Cousins. *"Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models with Application to Histograms"*. 2013. URL https://http://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf.